Census 2000 Experiment

April 17, 2003

Administrative Records Experiment in 2000 (AREX 2000) Outcomes Evaluation

FINAL REPORT

This research paper reports the results of research and analysis undertaken by the U.S. Census Bureau. It is part of a broad program, the Census 2000 Testing, Experimentation, and Evaluation (TXE) Program, designed to assess Census 2000 and to inform 2010 Census planning. Findings from the Census 2000 TXE Program reports are integrated into topic reports that provide context and background for broader interpretation of results.

Harley K. Heimovitz

Planning, Research, and Evaluation Division

USCENSUSBUREAU

Helping You Make Informed Decisions

Intentionally Blank

ACKNOWLEDGMENTS

The Administrative Records Experiment 2000 was conducted by the staff of Administrative Records Research at the U.S. Census Bureau, led by Charlene Leggieri. Questions and comments regarding this document can be directed to Harley K. Heimovitz at 301-457-4229.

Administrative Records Research Staff Members and Key Contributors to AREX 2000:

Bashir Ahmed	Mikhail Batkhan	Mark Bauder
Mike Berning	Harold Bobbitt	Barry Bye
Benita Dawson	Joseph Conklin	Kathy Conklin
Gary Chappell	Ralph Cook	Ann Daniele
Matt Falkenstein	Eleni Franklin	James Farber
Mark Gorsak	Harley Heimovitz	Fred Holloman
David Hilnbrand	Dave Hubble	Robert Jeffrey
Dean Judson	Norman Kaplan	Vickie Kee
Francina Kerr	Jeong Kim	Myoung Ouk Kim
Charlene Leggieri	John Long	John Lukasiewicz
Mark Moran	Daniella Mungo	Esther Miller
Tamany Mulder	Nancy Osbourn	Arona Pistiner
Ron Prevost	Dean Resnick	Pamela Ricks
Paul Riley	Douglas Sater	Doug Scheffler
Kevin A. Shaw	Kevin M. Shaw	Larry Sink
Diane Simmons	Amy Symens-Smith	Cotty Smith
Herbert Thompson	Deborah Wagner	Phyllis Walton
Signe Wetrogan	David Word	Mary Untch

and

Members of the AREX 2000 Implementation Group

Intentionally Blank

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Intentionally Blank

EXECUTIVE SUMMARY

This evaluation assessed the strengths and weaknesses of administrative data as a supplement or substitute for Census population counts. It compared county and subcounty population counts derived from administrative records to Census 2000 results. The Administrative Records Experiment in 2000 enumerated the population in two test sites that included two Maryland and three Colorado counties. The five counties offered distinct challenges to the enumeration process. Top-down and Bottom-Up enumeration method results were compared to Census population counts. The Top-Down method 'validates' administrative records addresses and assigns household members to Census blocks using Topographically Integrated Geographic Encoding and Referencing (TIGER) data. The Bottom-Up method is more stringent and required that each administrative records address match the Census Master Address File. Differences were presumed to vary by race, Hispanic origin, age, sex, imputation rates, and block-level characteristics, including vacancy and tenure rates. The results confirm that administrative records provide good estimates of Census counts at larger geographies. Some of the key findings include:

- **Administrative records provided county-level population counts that ranged from 97 to 102 percent of Census 2000 counts (using the Bottom-Up method).** And compared to Census 2000, more than 70 percent of tracts were within five percent, and 95 percent were within 25 percent of Census total population counts. But only 18 to 39 percent of *blocks* were within five percent of Census population counts. Age, race, sex, and Hispanic origin population counts produced worse results, due to multiple factors. The deficiencies were attributed to the files provided by federal agencies, their applicable dates, and administrative records processing operations. Each source of error can be minimized because of lessons learned through this evaluation process.

- **The Bottom-Up enumeration method produced more accurate household population counts for all counties.** The address-matching process was important because it validated addresses found in administrative records. This led to unmatched addresses being replaced by actual Census results. These activities were the most successful components in the administrative records processing operations. Several processes used for the Bottom-Up enumeration methodology were not evaluated in this report, including the request for physical address, clerical review, and field address verification. Request for physical address and clerical review provided a quality assurance check on the Bottom-Up results. The field address verification process relied on a small sample to develop correction measures and had little effect on the final results.

- **The youngest age group was consistently undercounted while the oldest age groups were overcounted.** Age under- and overcounting were attributed to demographic events, including birth, migration, and death, and the timeliness of reporting by agencies providing administrative records. This set of problems can be remedied by synchronizing file extracts from all participating agencies to coincide with an exact day, rather than time interval. However, age distributions are also affected by state policies in providing birth and death records, and Internal Revenue Service (IRS) 1040 and 1099 records that may have alternative address information that fails to place persons at their physical address.

- **Most of the race distributions did not accurately replicate Census results, which was attributed to weaknesses in the race imputation methodology.** Race imputation is perhaps the most deficient operation in the administrative records processing. For children, race information is seldom available because most federal agencies do not record these data. It is methodologically more difficult to impute race codes for adults as individuals or small areas (including tracts and blocks), compared to counties and states. However, combining administrative records sources and Census 2000 results will produce much better results than previously available.

These and other findings have led to the following key recommendations:

- **Identify and prioritize the goals, applications, and quality standards of administrative records processing.** This issue is important for focusing the work of a limited staff and providing assurances that objectives are successfully being met. Is tract or block-level accuracy more important and are there trade-offs that affect the accuracy of demographic characteristics? Should the immediate goal be accurate identification of individuals to improve linking with national surveys or would accurate tract-level characteristics be more useful? Should filing address be used when physical address cannot be identified? And what tolerance or level of error is acceptable for administrative records results? All of these conceptual issues should be addressed before further work commences.

- **Use the Bottom-Up enumeration method for subsequent administrative records processing and improve the master address file records.** Matching addresses between administrative records and the Census Master Address File provided significantly better results. The Geography Division will be enhancing the Master Address File, following Census 2000 results, and Bottom-Up estimates should also improve. However, there needs to be further research on non-city-style addresses and how to identify corresponding physical addresses. Improved address selection processing can achieve some success, but there is a need for additional research on address-related issues. This evaluation focused on the household population and special efforts need to be developed to enumerate group quarters.

- **Obtain file extracts from participating federal agencies that best reflect a particular date or narrow time period.** Inaccurate age distributions are primarily due to reporting lag or synchronicity between administrative files. First, data processing was based on files that were collectively current for Spring, 1999 or December, 1998, but compared to Census 2000. The direct consequence of this potential 15-month interval is that persons who died were reported in the administrative records, but not Census, while new births were reported in Census but not administrative records. This issue has a similar effect on movers and population mobility. Poor synchronization between federal files also impacts address selection processes because some files will have the most recent accurate information and others may not. Finally, race and Hispanic origin distributions may be indirectly affected because births were poorly enumerated and migrants tend to be minorities with higher fertility rates.

Additional efforts need to be focused on race imputation and children. The race imputation methods did not perform well. There is an immediate need for a new race imputation methodology that does not rely on model-based methods and accurately imputes race and ethnicity for tracts and blocks. Race and ethnicity generally come from Social Security files that fail to document this information in recent birth certificates. Additional data sources must be obtained, possibly through school enrollment data. Accurate demographic characteristics of parents may carry over to children and resolve many of these missing race identifiers. But there are problems using parent information for children. Divorced and separated couples with dependent children may have less accurate parent information and could be placed at one physical address rather than another.

Intentionally Blank

1. BACKGROUND

1.1 Introduction

The Administrative Records Experiment in 2000 (AREX 2000) was an experiment in two areas of the country designed to learn about the feasibility of an administrative records census (ARC) and the use of administrative records to enhance conventional decennial census processes. The first experiment of its kind, AREX 2000 was part of the Census 2000 Testing, Experimentation, and Evaluation Program. The focus of the program was to measure the effectiveness of new techniques and methodologies for decennial census enumeration. The test results lead to recommendations for further experiments and ultimately the design of the next decennial census.

Interest in an administrative records census dates back to a proposal by Alvey and Scheuren (1982), where Internal Revenue Service (IRS) records along with those from other agencies would form the core of an administrative record census. Knott (1991) identified two basic ARC models: (1) the Top-Down model that assembles administrative records from a number of sources, unduplicates them, assigns geographic codes, and counts the results; and (2) the Bottom-Up model that matches administrative records to a master address file, fills the addresses with individuals, resolves inconsistencies address by address, and counts the results. There have been a number of other calls for ARC research--see for example Myrskyla, 1991; Myrsklya, Taeuber and Knott, 1996; Czajka, Moreno and Shirm, 1997; Bye, 1997. All of the proposals fit either the Top-Down or Bottom-Up model. Knott also suggested a composite Top-Down/Bottom-Up model. Administrative records would be unduplicated using the Social Security Number (SSN), matched to the address file, and then proceed as in the Bottom-Up approach. In overall concept, AREX 2000 most closely resembles this composite approach.

More recently, direct use of administrative records in support of decennial applications was cited in several proposals during the Census 2000 debates on sampling for Nonresponse Followup (NRFU). The proposals ranged from direct substitution of administrative data for non-responding households (Zanutto, 1996; Zanutto and Zaslavsky, 1996; 1997; 2001), to augmenting the Master Address File development process with U.S. Postal Service address lists (Edmonston and Schultze, 1995:103). AREX 2000 provided the opportunity to explore the possibility of NRFU support.

The Administrative Records Research Staff (ARR) of the Planning, Research and Evaluation Division (PRED) performed the majority of coordination, design, file handling, and certain field operations of the experiment. They were supported by various other divisions within the Census Bureau, including Field Division, Decennial Systems and Contracts Management Office (DSCMO), Population Division, and Geography Division.

Throughout this report, rather than identifying individual workgroups or teams, we shall refer to the operational decisions made in support of AREX 2000 to be those of ARR; that is, we shall say that 'ARR decided to…' whenever a key operational decision is described, even though, of course, ARR staff were not the only decision makers.

1.2 Administrative Record Census-Definition and Requirements

In AREX 2000, an administrative record census was defined as a process that relies primarily on administrative records to produce the population content of the decennial census short form, with a strong focus on apportionment and redistricting requirements. Title 13, United States Code, directs the Census Bureau to provide state population counts to the President for the apportionment of Congressional seats within nine months of Census Day. In addition to total population counts by state, the decennial census must provide counts of the voting age (18 and over) population by race and Hispanic origin for small geographic areas, currently in the form of Census blocks, described in PL 94-171 (1975) and the Voting Rights Act (1964). These data are used to construct and evaluate state and local legislative districts.

AREX 2000 provided date of birth, race, Hispanic origin, and sex, although the latter is not required for apportionment or redistricting purposes. Geographically, AREX 2000 operated at the level of basic street address and corresponding Census block code. Unit numbers for multi-unit dwellings were used in certain address matching operations and one of the evaluations. But, household and family composition were not captured. AREX 2000 did not provide for the collection of sample long form population or housing data, needs that may be met by the American Community Survey (ACS) program. The design did assume the existence of a Master Address File and geographic coding capability similar to that available for Census 2000.

1.3 AREX 2000 Objectives

The principal objectives of AREX 2000 were twofold. The first objective was to develop and compare two methods for conducting an administrative records census, one that used only administrative records and a second that added some conventional support to the process in order to complete the enumeration. The evaluation of the results also included a comparison to Census 2000 results in the experimental sites.

The second objective was to test the potential use of administrative records data for some part of the Nonresponse Followup universe, or for the unclassified universe. Addresses that fall into the unclassified status have very limited information on them—so limited, in fact, that the occupancy status of some addresses must be imputed, and, conditional on being imputed "occupied", the entire household, including characteristics, must be imputed. In order to effectively use administrative records databases for substitution purposes, one must determine the type of households that are most likely to yield similar demographic distributions to their corresponding census households.

Other objectives of AREX 2000 included the collection of relevant information to support ongoing research and planning for administrative records use in the 2010 Census, and the comparison of an administrative records census to other potential 2010 methodologies. The results of these evaluations will assist in planning future decennial censuses, particularly those where administrative records are a primary source of data.

1.4 AREX 2000 Top-down and Bottom-up Methods

1.4.1 Top-Down

The AREX 2000 enumeration was accomplished with a two-phase process. The first phase involved the assembly and computer geocoding of records from a number of national administrative record systems, and unduplication of individuals within the combined systems. This was followed by two attempts to obtain and code physical addresses (clerical geocoding and request for physical address) for those that could not be geocoded by computer. Finally, there is a selection of "best" demographic characteristics for each individual and "best" street address within the experimental sites. Much of the computer processing for this phase was performed as part of the Statistical Administrative Records System (StARS), conducted in 1999 (Judson, 1999). As such, StARS 1999 was an integral part of the AREX 2000 design.

One can think about the results of the Top-Down process in two ways. First, counting the population at this point results in an administrative-records-only census. That is, the enumeration includes only those individuals found in the administrative records, and there is no other support for the census outside of activities related to geocoding. AREX 2000 provides population counts from the Top-Down phase so that the efficacy of an administrative-records-only census can be assessed. However, without a national population register as its base, one might expect an enumeration that used only administrative records to be substantially incomplete. And so a second way to think about the Top-Down process is as a substitute for an initial mail-out in the context of a more conventional census that would include additional support for the enumeration.

1.4.2 Bottom-up

The fundamental difference between the Bottom-Up and the Top-Down methods is that the Bottom-Up method matches administrative records addresses to a separately developed 'frame' of addresses, and based on this match, performs additional operations. In this experiment, an extract of the Census Bureau's Master Address File (MAF) served as the frame.[1]

The second phase of the AREX 2000 design was to complete the administrative-records-only enumeration by correcting errors in administrative records addresses through address verification (a coverage improvement analogue), and adding persons missed in the administrative records (a non-response follow-up analogue). This phase matched the addresses found in the Top-Down process to the MAF in order to assess their validity and to identify MAF addresses not matching administrative records addresses. A field address review (FAV) was used to verify unmatched administrative records addresses, and invalid administrative records addresses were excluded from the Bottom-Up selection of best address. Non-matched MAF addresses were canvassed in order to enumerate persons not found in the administrative records addresses. In AREX 2000, the canvassing process was simulated by adding persons found in unmatched Census 2000 addresses to adjusted administrative-records-only counts, thus completing the enumeration. Doing AREX 2000 as part of Census 2000 obviated the need to mount a separate field operation to canvass the unmatched MAF addresses. Considering the Top-down and Bottom-up processes

[1] In this report, we use the term 'MAF' generally. Our operations were based on extracts of the Decennial Master Address File (DMAF).

as part of one overall design, AREX 2000 can be thought of as a prototype for a more or less conventional census with the initial mailout replaced by a Top-Down administrative records enumeration. Figure 1.4 provides a conceptual overview of AREX 2000.

Note: The graphical description presented here is intended to convey the concept of both AREX methods when viewed in terms of the Bottom-up method as a follow-on process to the Top-down method.

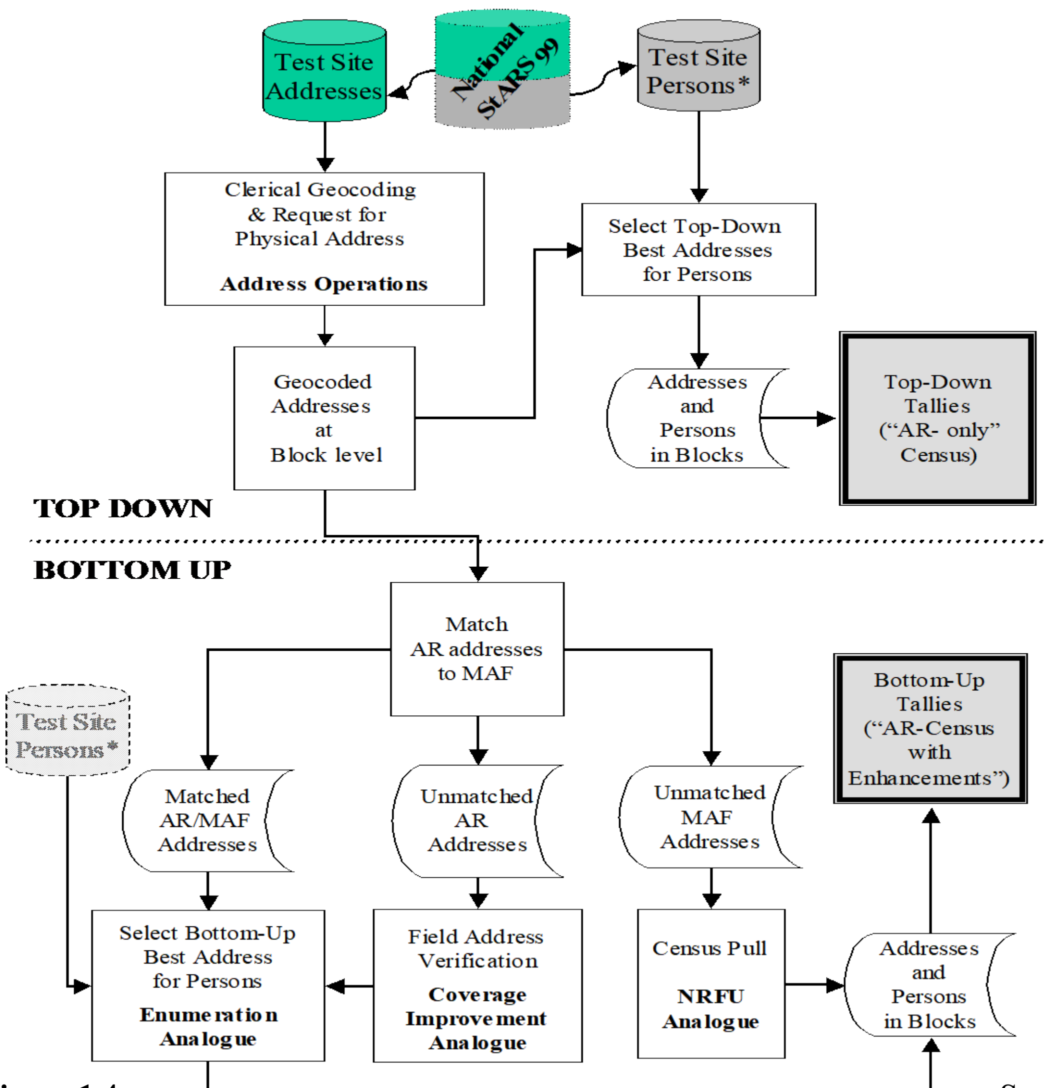


Figure 1.4. Summary Diagram of AREX 2000 Design

1.5 Experimental Sites

The AREX 2000 sample includes geographic areas that include both difficult and easy to enumerate populations (see table 1.5a). Two sites were selected that total approximately one million housing units and a population of approximately two million persons. One site included Baltimore City and Baltimore County, Maryland (MD). The other site included Douglas, El Paso and Jefferson Counties, Colorado (CO). The sites provided a mix of characteristics needed to assess the difficulties that might arise in conducting an administrative records census. Approximately half of the test housing units were selected based on criteria assumed to be easy-to-capture in an administrative records census and the other half were selected based on criteria assumed to be hard to capture. For example, areas having a preponderance of city-style addresses, single-family housing units, and older and less mobile populations were considered easier to enumerate. Demographic characteristics of the sites are given in Table 1.5b:

Table 1.5a Criteria for Selecting AREX 2000 Test Sites

Criteria	Easy-to-Capture	Hard-to-Capture
Address Type	City-style addresses with house numbers and street names	Non-city style address
Housing Unit Type	Single-family housing units	Multi-unit housing (rentals)
Age Category	Older age cohorts (65+)	Younger age cohorts (children less than 18 years old)
Population Type	Non-mobile population	Mobile population such as mobile homes occupants, immigrants, movers
Race	White and Black population	Populations not dominated by Whites or Blacks

Table 1.5b: Key Demographic Characteristics of the AREX 2000 Sites

	Baltimore County, MD	Baltimore City, MD	Douglas County, CO	El Paso County, CO	Jefferson County, CO	United States
Total Population[1]	754,292	651,154	175,766	516,929	527,056	281,421,906
White[1]	74.4%	31.6%	92.8%	81.2%	90.6%	75.1%
Black[1]	20.1%	64.3%	1.0%	6.5%	0.9%	12.3%
American Indian, Eskimo, or Aleut[1]	0.3%	0.3%	0.4%	0.9%	0.8%	0.9%
Asian or Pacific Islander[1]	3.2%	1.5%	2.6%	2.7%	2.4%	3.7%
Other Race[1]	0.6%	0.7%	1.4%	4.7%	3.2%	5.5%
Multi-Race[1]	1.4%	1.5%	1.9%	3.9%	2.2%	2.4%
Hispanic[1]	1.8%	1.7%	5.1%	11.3%	10.0%	12.5%
Median age[1]	37.7 yrs	35.0 yrs	33.7 yrs	33.0 yrs	36.8 yrs	35.3 yrs
Crude Birth Rate[2]	12.6	14.9	19.0	15.7	12.5	14.9[3]
Crude Death Rate[2]	9.9	13.1	2.7	5.5	6.0	8.6[3]
1990-2000 Change[4]	9.0%	-11.5%	191.0%	30.2%	20.2%	13.2%

Note: all values include household and group quarters residents
[1] Census 2000 results
[2] 1998 rates per 1000; from MD Dept. of Health and Mental Hygiene and CO Dept. of Public Health and Environment
[3] 1998 rates per 1000; from www.fedstats.gov
[4] Census 1990, 2000 results

1.6 AREX 2000 Source Files

The administrative records for AREX 2000 were drawn from the StARS 1999 data base. There were six source files with national coverage selected for inclusion in StARS The files were chosen to provide the broadest possible coverage of the U.S. population and compensate for the weaknesses or lack of coverage of a given segment of the population inherent in any one source file. At a minimum, the files had to have for each record, a name, Social Security Number (SSN), and street address.

The national level files that contributed to the StARS 1999 database and to AREX 2000, were:

- Internal Revenue Service (IRS) Tax Year 1998 Individual Master File (1040).
- IRS Tax Year 1998 Information Returns File (W-2 / 1099).
- Department of Housing and Urban Development (HUD) 1999 Tenant Rental Assistance Certification System (TRACS) File.
- Center for Medicare and Medicaid Services (CMS) 1999 Medicare Enrollment Database (MEDB) File.
- Indian Health Services (IHS) 1999 Patient Registration System File.
- Selective Service System (SSS) 1999 Registration File.

Table 1.6 displays the primary reason each file was included in the StARS database and the approximate number of input records associated with each.

Table 1.6: AREX 2000 Source File Characteristics

File	Targeted Population Segment	Address Records	Person Records
IRS 1040	Taxpayer and other members of the reporting unit with current address	120 million	243 million
IRS W2/1099	Persons with taxable income who might not have filed tax returns	598 million	556 million
HUD TRACS	Low income housing population (possible non-taxpayers)	3 million	3 million
Medicare File	Elderly population (possible non-taxpayers)	57 million	57 million
IHS File	Native American population (possible non-taxpayers)	3 million	3 million
SSS File	Young male population (possible non-taxpayers)	14 million	13 million
	Total	**795 million**	**875 million**

Notes: The variance between the number of address records and person records within the input source files is a result of the following source file characteristics:

1. The number of address records column is generally synonymous with the total record count on the input file.
2. Each IRS 1040 input record may reflect up to six persons (primary filer, secondary, and dependents).
3. Each SSS input record may reflect two addresses - defined as current and/or permanent address.
4. The IRS W-2/1099 file undergoes a preliminary unduplication and clean-up process prior to the initial file edit process.

1.6.1 Timing

An important limitation of AREX 2000 is the gap between the reference period for data contained in each source file and the point-in-time reference of April 1, 2000 for the Census. The time lag has an impact on both population coverage--births, deaths, immigration and emigration--and geographic location--housing extant, and geographic mobility. As an example, both IRS files include data for tax year 1998 with an expected current address as of tax filing time close to April 15, 1999. But the IRS 1040 file identifies persons in the tax unit as of December 31, 1998. Table 1.6.1 shows the reference periods of the files, which generally have a cutoff date one year prior to the enumeration of Census 2000.

Table 1.6.1: AREX 2000 Source File Reference Dates

Source File	Cut-off Date	Requested Cut Date	Universe
Indian Health Svc.	04/01/99	04/01/99	All persons alive at cut-off date
Selective Service	Note 2	04/01/99	Males between the age of 18 - 25[2]
HUD TRACS	04/01/99	04/01/99	All persons on file as of cut-off date
Medicare	Note 3	04/01/99	All persons alive at cut-off date
IRS 1040	12/98	09/30/99[1]	Individual tax returns for tax year 1998
IRS W-2 / 1099	12/98	04/01/99	Forms W-2 and all 1099 forms tax year 1998

1. File Cut date is for posting cycle weeks 1-39 only for IRS 1040, and weeks 1-41 for IRS 1099 files. Weeks 40-52 (and 42-52 respectively) were not included in StARS '99. This file reflects the most current address on file for the taxpayer. It could be an address that has been updated since the 1998 tax return was posted.
2. Cut-off date is same as dates used to define universe: persons born after April 2, 1972 and on (or before) April 1, 1980.
3. Universe also defined as persons with a death date of 12/31/1989 or later.

1.6.2 State, Local, and Commercial Files

ARR staff decided not to use state and local files and commercially available databases in the AREX 2000 experiment.[2] Statistical evidence is limited, but various reports from ARR staff indicated that state and local files exist in an extremely diverse variety of forms, with equally diverse record layouts and content (for historical information, see Sweet, 1997; Buser, Huang, Kim, and Marquis, 1998; and other papers in the Administrative Records Memorandum Series). Furthermore, ARR staff reported that it was quite time-consuming and intricate to develop the interagency contractual arrangements necessary to use state and local files. Public opinion results such as Singer and Miller (1992), Aguirre International (1995), and Gellman (1997), convinced ARR staff that public sensitivity to the idea of linking commercial databases with government databases (other than for address processing) would be too great, and that such a linkage would be unwise. The American Business Index (or ABI) file was used to identify addresses that were commercial rather than residential, and a Group One product, Code One, used to standardize addresses.

In addition to acquisition and processing difficulties, consideration of the use of state and local files raises an equity issue in a decennial census context. Since it is not possible to obtain an

[2] Such as state and local tax returns, drivers license files, local utilities, assessor's records. Commercially available databases include direct mailing lists, credit card databases, etc.

exact count of the population in its entirety, public perception of fair treatment in the decennial census process is important. This means that the accuracy of the counts must be seen as uniform between and within states. The use of data from only certain states or localities would compromise notions that decennial census methods must treat all parts of the country equitably.

1.6.3 Census Numident

Census Numident was critical in the creation of the StARS database, and a source of most of the demographic characteristics and some of the death data. Census Numident was created by ARR primarily to validate Social Security Numbers (SSNs) used in the administrative records and to substitute demographic variables missing from source files. The Census Numident is an edited version of the Social Security Administration's (SSA) Numerical Identification (Numident) File. The SSA Numident file is a numerically ordered master file of assigned Social Security Numbers (SSN) that has up to 300 entries for each SSN record, though most SSNs have two records. Each entry represents an initial application for a SSN or an addition or change (referred to as a transaction) to the information pertaining to a given SSN. The SSA Numident contains all transactions (and therefore, multiple entries) ever recorded against a single SSN. The SSA Numident available for StARS 1999 reflected all transactions through December 1998.

The Census Numident was designed to collapse the SSA Numident entries to reflect "one best record" for each SSN containing the 'best' demographic data for each SSN on the file. However, all variations in name (including married names, maiden names, nicknames, etc.) and date of birth were retained as part of the Census Numident, as Alternate Name Date of Birth Files, respectively. For the Census Numident, selection criteria were established for each Census 2000 Short Form demographic variable (after minor edits were accomplished in an effort to standardize the variables). The short form variables include date of birth fields, gender, race, and Hispanic origin. Following the edit, unduplication, and selection processes, the SSA Numident file was reduced from 677 million records to about 396 million records that comprise the Census Numident file.

1.6.4 What Effect did Race Imputation have on the AREX 2000 Counts?

AREX 2000 Evaluation Outcomes focuses on single race reports and compares Census single race responses to equivalent AREX race categories. However, the Office of Management and Budget (OMB) revised the classification of race and ethnicity categories in 1997, and Census 2000 includes multiple race and 'some other race' (SOR) reports. AREX race assignment relied on a complex decision-making process that addressed the reliability of AREX source records, their frequency of occurrence, and a statistical estimation methodology for calculating race probabilities. But the complications in assigning race go beyond the logical and mechanical processes of determining the most accurate race of an individual. Many federal agencies do not collect race information, have different race classifications, or changed their classification categories over time. Reconciling these differences in statistical decision models invites errors that cannot be avoided. But some of this decision-making process may require inferences from large numbers of individuals onto smaller groups. The result of such inferences produces its own set of errors, because applying aggregate results to individuals invites 'regression towards the mean.' That is, a best guess is based on an average that may not fit many of those individuals. The consequence is that larger aggregate measures, for example counties, may have reliable estimates, while tracts and blocks have increasingly greater error rates.

1.6.5 What Effect did Vital Events have on the AREX 2000 Counts?

The five test counties have some striking demographic differences between them, despite the different enumeration criteria (Tables 1.5a, 1.5b). The Census population is counted on the same day for all households, but administrative records counts may measure the same items on different days. This is due to the cycle of events leading to the recorded administrative record, and is affected by the type of demographic process, the intermediate agencies that process that data and their processing dates, and whether ARR has received the most recent data extracts. With the vital events of birth and death, the event is first recorded by county and state agencies before reporting to federal agencies that collect administrative data. Delays in the reporting process can affect the reliability of administrative records. The mortality rate is quite high in Baltimore City, while Douglas County has a high birthrate and low mortality rate. These rate differences may affect population counts for persons aged 0-4, as well as older persons who have higher mortality rates. Inaccurate age counts for the oldest and youngest persons may also have an indirect effect on race and Hispanic origin counts. For example, if most births in Douglas County are in Hispanic families and the 0-4 age group counts are unreliable, then Hispanic counts may be affected at block, tract, and potentially county levels. Alternatively, IRS records may not cover all persons because they are non-filers, while late-filers may have been excluded from some extracts.

Population change between 1990 and 2000 is also a consideration and the AREX counties have some key differences. Baltimore City and County had lower growth rates than the three CO counties and the U.S. national average. There is no explicit means of recording migration in administrative records. Migration is captured by address changes that are dependent upon the type of participant and their active involvement in that federal program. Delayed or lagged reporting is likely to affect each of the five counties in different ways and especially at block- and tract-levels. But migration may consist of new inter-regional migrants from other areas of the U.S., as well as intra-regional migrants. There is some evidence of intra-regional migration from Baltimore City to Baltimore County, while the CO counties have grown through inter-regional migration. Migrating Baltimore City residents may have settled in suburban Baltimore County, while migration from other U.S. cities, Central America, and Asia fueled the rapid population expansion in CO. The key issue in these two types of migration is that the population composition of inter-regional and intra-regional migrants is likely to differ.

1.7 AREX 2000 Evaluations

Currently, four evaluations are being completed.

The **Process Evaluation** documents and analyzes selected components or processes of the top down and bottom up methods in order to identify errors or deficiencies. It is designed to catalog the various processes by which raw administrative data became final AREX counts and attempts to identify the relative contributions of these various processes.

The **Outcomes Evaluation** is a comparison of top down and bottom up AREX counts by county, tract, and block level counts of the total population by race, Hispanic origin, age groups and gender, with comparable decennial census counts. This evaluation is outcome rather than process oriented.

The **Household Evaluation** assesses outcomes of the Bottom-Up method, the potential for nonresponse substitution and unclassified imputations, and predictive capability. Nonresponse Followup substitution assesses the feasibility of using administrative records, in lieu of a field interview, to obtain data on non-responding census addresses via the bottom up method.

The **Request for Physical Address Evaluation assesses the impact of non-city-style addresses.** These addresses present a significant hurdle to the use of an administrative records census on either a supplemental or substitution basis is the determination of residential addresses and their associated geographic block level allocation for individuals whose administrative record address is a P.O. Box or Rural Route. AREX 2000 tested a possible solution in the form of the Request for Physical Address operation. Several thousand letters were mailed to P.O. Box and Rural Route addresses requesting the receiver to reply with their residential address for purposes of block level geocoding. This report documents in detail the planning and implementation of the operation. It also analyzes the results of the operation and assesses its potential future use as part of an ARC.

2. METHODOLOGY

2.1 Conceptual Design

This evaluation incorporates a variety of methods to accomplish its objectives, including univariate and multivariate statistical analyses of AREX-Census differences, and spatial/ecological maps that examine the distributions of key measures. AREX 2000 Outcomes tries to disentangle the influence of demographic change and AREX processing, coverage and data quality issues, while trying to answer the general question:

What factors influenced the accuracy of the AREX county and subcounty results, what actions could improve the quality and coverage of administrative records, and what are the limitations of administrative records as a reliable source of intercensal population counts?

AREX 2000 Outcomes measures how well AREX simulates Census 2000 results at county and subcounty levels and identifies weaknesses in AREX processing. Key demographic characteristics are assessed, as well as differences between Bottom-Up and Top-Down results. A series of research questions provides a conceptual outline of the basic elements of the evaluation. General questions at larger geographies are posed first:

Q1: How well does AREX measure total Census population at the county level, and how do the results differ by whether the Top-Down or Bottom-Up sample was used?

Q2: How do county-level differences between AREX and Census differ by age, race, sex, and ethnicity, as well as Top-Down and Bottom-Up differences?

The AREX and Census voting age population counts are compared for voting districts. This comparison provides a rough measure of how well administrative records data could provide redistricting information. The county-level analyses are then repeated for tract and block-level comparisons. Greater differences between AREX and Census counts are more likely at smaller geographies. But focusing on smaller geographies allows more detailed analyses of general neighborhood characteristics (e.g., block-level race/ethnic composition) and whether these attributes are linked with AREX-Census differences:

Q3: How well does AREX measure the voting age population (age 18+) of state legislative districts?

Q4: How well does AREX measure tract-level total, age, race, sex, and Hispanic origin counts, as well as block-level totals?

Q5: How did AREX processing and imputation of race codes impact the county and sub-county race distributions?

Q6: What are the most important spatial/geographic issues in comparing AREX-Census demographic characteristics?

The third and final stage of the evaluation is the most detailed and includes a multivariate analysis of AREX-Census differences with thematic map analyses:

Q7: What are the key predictors of AREX-Census differences using multivariate regression models and how well do these predictors estimate AREX-Census errors?

Q8: What is the spatial/geographic distribution of AREX-Census regression residuals and what unobserved/unmeasured spatial relationships exist in the results?

2.2 Variable Constructs and Measures

The terms 'undercount' and 'overcount' are used to describe how well AREX counts match Census results and have no further connotation. That is, undercounts and overcounts reflect any of several problems, including coverage issues, coding, and processing errors. Outcome and predictor constructs are distinguished and used to highlight AREX-Census and Bottom-Up and Top-Down differences. Variable definitions used in this evaluation include:

Difference

The simple difference between AREX and Census gauges the county-level over and under-counts:

$$DIFF(A_i,C_i) = A_i - C_i$$

where: A_i = AREX tallies in county

C_i = Census 2000 tallies in county

Algebraic percent error (ALPE)

Smaller geographies vary by population size, which can be used to standardize AREX-Census differences. AREX and Census counts are the inputs for calculating the algebraic percent difference (or, where one is taken as the standard, the algebraic percent error), for the i^{th} county, tract, or block:

$$ALPE(A_i, C_i) = \sum_{i=1}^{j} \frac{A_i - C_i}{C_i}$$

where: A_i = AREX tallies in the i^{th} county, tract, or block; and

C_i = Census 2000 tallies in the i^{th} county, tract, or block

There are two problems when computing ALPEs: zero blocks and inflated ALPEs. Zero blocks occur when AREX reports at least one person having a particular characteristic but Census does not. Because Census is being used as the standard and is the denominator for the ALPE measure, these zero blocks are undefined. For the purpose of block comparisons, these zero blocks will be omitted from the analyses. However, county and tract-level counts and comparisons include these blocks because they are aggregated at larger geographies. The regression residual analyses describe the spatial distribution of zero-blocks in the AREX sites.

Inflated ALPEs occur because some blocks have very small denominators that tend to produce large ALPEs, despite small differences between AREX and Census. That is, blocks and tracts with smaller populations are more apt to have larger ALPEs. There are several ways to address this issue. Median block values can be used so that inflation can be minimized. However, inflation will still be present and the use of medians provides less information about distributional characteristics. A second alternative is to trim or topcode large values before calculating site-level means. This alternative sets all values greater than the 95th percentile at the 95th percentile. A third alternative is to apply weights to all block or tract-level observations that equalize the impact of observations on aggregate measures. For example, blocks with small population counts may have larger ALPEs but have the same influence as large populations on computed means. Blocks and tracts with high population densities have a greater influence on means, while low density, rural blocks or tracts have a smaller influence. All of the three approaches are imperfect, but applying the second alternative for both tract and block ALPEs provides a less-biased estimate of AREX-Census differences.

Shannon-Wiener Index of Diversity

This measure is widely used for estimating the biodiversity of plant and animal species within specified land areas (Krebs, 1989). It provides a concise index of the county, tract, and block-level racial/ethnic composition and can be used to calculate separate AREX and Census measures.

$$H(p) = -\sum p_k \log_e(p_k)$$

where: p_k = race/Hispanic proportion in the kth category

Index of Dissimilarity

This measure combines the features of ALPE and Shannon-Wiener to calculate race/Hispanic and age indices (Shryock and Siegel. 1973):

$$D(A_j, C_j) = \frac{1}{2}\sum_{j=1}^{r}\left|\frac{A_{ij}}{\sum_{j=1}^{r} A_{ij}} - \frac{C_{ij}}{\sum_{j=1}^{r} C_{ij}}\right|.$$

where: i = age or race subgroup of the jth county, tract, or block;

A_{ij} = AREX tallies in the jth county, tract, or block; and

C_{ij} = Census 2000 tallies in the jth county, tract, or block

Race

Both AREX and Census versions of this variable use reported single race values with categories

White, Black, American Indian (AI), and Asian-Pacific Islander (API). The Hispanic origin of the race categories is ignored. A small proportion of respondents self-reported multiple races in their Census forms. Limited analyses will examine the influence of multi-race reporting on under- and overcounts of AREX results. Race is used in calculating differences, ALPEs, Shannon-Wiener and Dissimilarity Indices.

Hispanic origin

Both AREX and Census versions of this variable use reported single Hispanic origin values (yes/no) and ignore race category. Hispanic origin is used in calculating differences, ALPEs, Shannon-Wiener and Dissimilarity Indices.

Sex

Used for calculating differences and ALPEs with male and female categories.

Population density

Population density for blocks and tracts is calculated using Census total population values. Quintile groups of increasing population density are used in the multivariate analyses. The cutpoints of these quintiles were obtained from the combined MD and CO blocks. This allows the same quintile cutpoints to be used in both AREX sites and facilitates comparisons between the sites. Because the CO AREX site is more rural, CO blocks and tracts have smaller population densities, compared to the MD AREX blocks and tracts.

Population density = total population of block/tract i / block/tract area of i (in square miles)

Neighborhood characteristics

Neighborhood categories are distinct for the MD and CO AREX sites. Neighborhood categories are defined from factor analysis results of block characteristics that distinguish differences in population demographics, population density, and the availability and type of housing.

Vacancy rate

Vacancy rate uses Census-reported values of housing unit vacancies within blocks and tracts. Models use a binary measure of greater than/less than median vacancy rate.

Vacancy rate = vacant housing units in tract or block i / total housing units in tract or block i

Rental rate

Rental rate uses Census-reported values of rented units within blocks and tracts. Models use a binary measure of greater than/less than median rental rate.

Rental rate =occupied rental units in tract or block i / total housing units in tract or block i

Presence of non-relative household members

Census-reported number of housing units with non-relative household members.

Non-relative rate = housing units with non-relative household members in tract or block i / total housing units in tract or block i

Multi-race reporting on Census 2000

Number of persons claiming multiple races on Census forms. Models use a binary flag indicating the presence/absence of multi-race reports by individuals.

Multi-race rate = individuals claiming multi-race in tract or block i/ total persons in tract or block i

2.3 Analysis Plan

The analysis plan has four segments of increasing complexity that provide summary, bivariate, spatial/ecological, and multivariate analyses that control for compositional differences between counties. A brief description of the goals and types of analyses in these categories is shown below:

Summary analyses

This section is intended to be a top-level, descriptive summary of AREX-Census differences, by county, voting district, and tract. County-level counts and proportions are compared and display the raw, untransformed numbers not shown in the detailed analyses. The count differences describe the aggregate under- and over-counts of age, race, sex, Hispanic categories, while the proportions show the contribution the age, race/ethnicity, sex categories have on the under- and overcounts. Analyses of voting districts, tracts, and blocks emphasize the distributional aspects of AREX-Census differences.

Bivariate analyses

A second set of bivariate analyses examines how the AREX race assignment and imputation methodology affected the race ALPE results. County, tract, and block ALPEs are analyzed by key race decision flag indicators, including the proportion of persons with imputed and non-imputed race variables.

Spatial/ecological maps

One important aspect of the bivariate analyses is the ecological variation between blocks and tracts. Thematic maps profile the heterogeneous nature of each AREX site and the spatial distribution of key housing and population characteristics of the MD and CO tracts. The map profiles and bivariate results are then used to focus on the spatial aspects of key bivariate relationships. The profile maps include:

- Vacant housing units.
- Population density.
- Shannon-Wiener index of diversity for age and race/ethnicity.

Additional maps that describe key AREX-Census tract differences by AREX site include:

- Index of dissimilarity for age groups.
- Index of dissimilarity for race/ethnicity.
- AREX-Census ALPEs for total population.
- AREX-Census ALPEs for persons age 0-4.
- AREX-Census ALPEs for persons aged 65+.
- AREX-Census ALPEs for Blacks.
- AREX-Census ALPEs for Hispanics.

Multivariate analyses

This section builds on the results of the bivariate analyses to develop predictive models of AREX-Census differences. The block-level, multivariate analyses consider the qualitative characteristics of neighborhoods, which are hypothesized to be more or less stable, based on the composition and dynamics of households. Factor analysis is used to distinguish types of neighborhoods within each AREX site. The neighborhood groupings are mutually exclusive categories and can be summarized as:

Maryland neighborhoods:

- Larger proportions of Blacks and younger persons (Black-younger).
- Predominantly older (55+) and White (White-older).
- Multiethnic neighborhoods with Blacks and Hispanics (multiethnic).
- Multi-ethnic neighborhoods with Asian-Pacific Islanders, Hispanics, and younger persons (multiethnic-younger).

Colorado neighborhoods:

- Multi-ethnic neighborhoods with Blacks, Asian-Pacific Islanders, and Hispanics (multiethnic).
- Predominantly renters aged 25-34 (young-renters).
- Older persons in suburban (moderate density) neighborhoods (older-suburban).
- Neighborhoods with higher mobility (vacancy) rates (transient-vacant).

AREX under- and over-counts are hypothesized as having distinct sets of predictors and are estimated in separate regression models of under- and overcounted blocks. The expected predictors of block-level ALPEs are shown below:

County indicator
+ Population density
+ Neighborhood characteristics
+ High vacancy rate
+ High rental rate
+ High non-relatives
+ White quintile groups
+ High race proportions (excluding Whites)
+ High Hispanic proportion
+ High age group proportions
+ AREX race imputation variables
+ Other AREX processing variables

The models emphasize Bottom-Up results and the possible causes of error affecting Blacks, Hispanics, and selected age groups. The selected analyses were based on results from the descriptive analyses that follow. Six models predicting block-level AREX-Census ALPEs are estimated with the following outcome and sample characteristics:

1) total population ALPEs using MD Bottom-Up sample.

2) total population ALPEs using CO Bottom-Up sample.

3) age 0-4 ALPEs using CO Bottom-Up sample.

4) age 85+ ALPEs using MD Bottom-Up sample.

5) Black ALPEs using MD Bottom-Up sample.

6) Hispanic ALPEs using CO Bottom-Up sample.

But the distribution of ALPEs is truncated at –1 (minus one) when the AREX population equal zero, and the small Census blocks have inflated overcounts. To compensate for this difficult to transform ALPE distribution, the values are categorized into groups. Grouping the ALPE values and creating subsamples helps reduce the difference in actual and predicted errors, or residuals. Each ALPE dependent variable is assigned to one of five subgroups for separate regression models, based on their interquartile ranges:

Group 1: greatest ALPE undercount reflecting 12.5 percent of blocks.

Group 2: next largest ALPEs reflecting 25 percent of blocks.

Group 3: next largest ALPEs reflecting 25 percent of blocks.

Group 4: next largest ALPEs reflecting 25 percent of blocks.

Group 5: next largest ALPEs (greatest overcounts) reflecting 12.5 percent of blocks.

For most of the dependent variables, groups one and two include undercounts while Groups four and five are overcounts. Generally, group three has the smallest ALPE scores (both under- and overcount) and includes the zero-score. Groups one and five are wider intervals and include the most extreme values, though outliers were previously topcoded to the 95th percentile.

A preliminary set of models compares the quartile group memberships in categorical regression models and includes all blocks with complete data. This set of models is useful for comparing the between-group differences in the blocks, providing indicators of how the demographic, imputation, and processing issues affected the accuracy of the AREX counts.

Standard, multivariate regression models are then estimated for each of the five groups using the narrowed ALPE interquartile ranges as dependent variables. Blocks with undefined ALPEs, where AREX counts exist but Census indicates no persons, are again excluded from the analyses. The regression parameter estimates are then used to calculate predicted values and residuals (actual block population ALPE – predicted block ALPE) for each of the blocks. The residuals are presented in thematic maps to highlight the ecological issues underlying AREX-Census deviations and unmeasured/ignored block-level heterogeneity. Block-level heterogeneity is potentially linked with the unobserved social characteristics of AREX-Census differences.

3. LIMITS

Study and Data Limitations

There are three potential sources of error that impact the AREX counts:

- errors in raw administrative records.
- inaccurate recording of demographic events.
- ARR processing decisions.

These influences can interact with the accuracy of total counts and age, race/ethnicity, sex distributions, and impact whether persons are correctly matched to their block or tract of residence. And while ARR processing decisions attempted to minimize and correct deficiencies in the raw administrative records, the resultant data could have been altered but *not* made more accurate after processing decisions were implemented. The main issues that affect these sources of error are summarized by main category.

Errors in raw administrative records

Three processed datasets are used in this evaluation, including Top-Down AREX counts, Bottom-Up AREX counts, and Census 2000 results. Top-down counts were obtained by processing administrative records to place persons within their block of residence. Bottom-up counts can be described as processed Top-Down counts that exclude group quarters residents, with edited address information and some imputed demographic measures. The Census records in this evaluation exclude group quarters residents and correspond to AREX counties, tracts, and blocks.[3] Some observed patterns in the AREX files include:

- Most administrative data have limited coverage or cover selected populations; for example Medicare records cover the 65+ population very well, while the Social Security Administration provides more accurate information for active participants, including employed persons and beneficiaries; most administrative records do not fully cover children and/or provide limited characteristics.
- Some administrative data provide information for all age categories but only in selected locations; for example, the Indian Health Service provides good information on its participants if they live within tribal areas or reservations, but provides no information for other locations.
- Definitions can vary between data sources, locations, and the time when data were collected; for example, race definitions have not only changed over time, but some agencies collect multiple race and ethnicity characteristics, while agencies in some locations may have unique circumstances (APIs in Hawaii and AIs near tribal areas may

[3] Bottom-up and Top-Down will be used to refer to these methods and as file names for the remainder of this report.

be treated differently); generally, the Social Security Administration provides the most consistent and complete coverage of the AREX population and U.S. residents.

- Under-reporting and non-coverage may occur if persons are not active participants with data collectors, especially persons at the lowest socioeconomic levels who may be unemployed or disabled, do not have interest-bearing bank accounts, and do not file tax returns.
- Raw administrative records are also likely to have different posting and processing dates, so that more recent demographic events may have better or worse coverage, depending upon the processing standards of the data-providing agency; the extent of reporting lag and differences across the raw administrative files have not been fully evaluated.

Inaccurate recording of demographic events

- New births are often not registered in administrative records because of a lagged response by data collectors; new birth data are generally identified through tax returns and Social Security records, though these sources do not fully disclose race and ethnicity measures.
- New deaths may also be subject to a lagged response by data collectors, which were identified by the Social Security Administration and the Center for Medicare and Medicaid Services, (formerly the Health Care Financing Administration); because death records are obtained from states having different disclosure rules and processing policies, there may be some geographic biases in the accuracy and timeliness of death records.
- Migration and mobility can be identified through tax returns, but addresses are likely to be updated on an annual basis and also be subject to a lagged response; updated address records for other data sources have varying accuracy and timeliness.

ARR processing decisions

- Decision rules were made by ARR to unduplicate and match all of the administrative records, and the resultant data may be biased by age, race/ethnicity, sex, and household address.
- Demographic imputation processes were implemented to select the best race/ethnicity identifiers and fill in missing age and sex characteristics; the resultant identifiers may also be biased and/or less accurate than desired.
- Missing and problematic address information that failed to match the Census Master Address File (MAF) underwent further processing; in some cases, persons at these addresses were statistically allocated to blocks based on in-person field address verification (FAV estimation); in other instances, the actual Census records were pulled in to replace these persons in the AREX data (Census pull).

4. RESULTS

4.1 Net Differences in AREX and Census Population Counts

Summary of results: AREX undercounted total household population in four of the five counties with algebraic percent errors of 97 to 102 percent of Bottom-Up Census results (Table 4.1). The Bottom-Up results were generally better than the Top-Down results: Bottom-Up had more stringent processing specifications and added 'Census pull' households for unmatched addresses (census pull rates are shown in Table 4.5). If AREX was unbiased and counted all demographic groups in the same way, we could expect undercounts for all demographic categories to have the same relative size. However, older persons aged 65+, especially persons aged 85+, and college-aged persons (aged 20-24) were overcounted. The second important finding is that Hispanics in MD and Blacks and APIs in CO were overcounted, but represent small minorities in those counties. Whites were overcounted in Baltimore City and County, where they reflect a smaller share of County population, compared to the CO counties. Demographic processes affect the accuracy of AREX counts in the youngest, oldest, and college-age age categories. The accuracy of race and Hispanic counts is subject to more complex operational, demographic, and administrative processes.

Table 4.1: Top-down and Bottom-up Counts of Total Household Population by County[1]

	Top-down Results				Bottom-up Results			
	AREX	Census	Difference	ALPE	AREX	Census	Difference	ALPE
Baltimore County	696,183	736,652	-40,469	-5.5%	728,205	736,652	-8,447	-1.1%
Baltimore City	570,648	625,401	-54,753	-8.8%	636,729	625,401	+11,328	+1.8%
Douglas County	148,270	175,300	-27,030	-15.4%	169,640	175,300	-5,660	-3.2%
El Paso County	456,891	501,533	-44,642	-8.9%	494,253	501,533	-7,280	-1.5%
Jefferson County	473,495	519,326	-45,831	-8.8%	503,622	519,326	-15,704	-3.0%

[1]AREX Top-Down counts include persons who were later identified in Bottom-Up as group quarters residents; Bottom-Up and Census counts exclude group quarters residents.

TOTAL POPULATION (see Table 4.1)

- AREX (Top-Down) undercounted all five counties with the greatest net undercounts in Baltimore City and El Paso and Jefferson Counties
- Bottom-up undercounts are much smaller than Top-Down undercounts in all five counties for total population and demographic characteristics; Bottom-Up showed the greatest improvements in Baltimore City and El Paso County

SEX

- Males and females are undercounted in all five counties (except Baltimore City); female undercounts are greater than male undercounts in all five counties for both methods; comparing Bottom-Up and Top-Down results, the differential undercount of females is smaller in CO than in the MD counties.

AGE

- In the MD counties, Top-Down overcounts the 75+ population and undercounts all other age groups; Bottom-Up overcounts the 20-24, 25-34, 35-44, and 65+ age groups and undercounts all other age groups.
- In the CO counties, generally, age 20-24 and 65+ age groups are overcounted and other age groups are undercounted for both Top-Down and Bottom-Up samples.
- In both MD and CO, Top-Down undercounts are greatest for the 0-19 age groups and show the greatest improvements for Bottom-Up counts.
- At the oldest ages in the MD counties (somewhat less in CO), the 85+ age group is overcounted, while 65-74 and 75-84 age groups are both undercounted and overcounted in Top-Down and Bottom-Up; given that mortality rates and increasing overcounts are associated with advancing age, the results suggest lagged reporting of deaths by agency administrators.

RACE

- In the MD counties, Hispanics were overcounted and other minority race groups were generally undercounted in Top-Down and Bottom-Up; Whites and Blacks were overcounted in Baltimore City (Bottom-Up) where Blacks are a majority of the population.
- In the CO counties, Blacks and APIs were generally overcounted while AIs and Hispanics were undercounted in Top-Down and Bottom-Up.

Some initial insights from the net under- and overcounts are evident in Table 1.5b (section 1.5). One general pattern is the relationship between share of minority population and under- and overcount. Hispanics are a smaller proportion in the MD counties and have larger undercounts. Similarly, Whites in Baltimore City and Blacks and APIs in the CO counties were both overcounted. But Hispanics in MD and Blacks and APIs had higher rates of imputation from general[4] and tax form methods. That is, the AREX-Census differentials were larger because under- and overcounts have smaller population bases (compared to the majority race group) and higher potential error rates (from imputation). Differences between Bottom-Up and Top-Down results are likely due to the address-matching requirement of Bottom-Up that eliminates potentially inaccurate records, and the added Census pull records, which directly affect the AREX-Census comparisons. Census pull rates were large for Baltimore City and Douglas County, but both counties also experienced significant population change between 1990 and 2000.

4.1.1 AREX-Census Algebraic Percent Errors

Summary of results: The county-level ALPE results provide a simple display of aggregate results and suggest how analyses of smaller geographies are likely to be affected by administrative reporting delays, the impact of the race imputation model, and differences between Top-Down and Bottom-Up methods. Generally, Bottom-Up under- and overcounts were smaller for race, Hispanic origin, age, and sex groups. Males and females are undercounted in four of five counties, with female undercounts slightly greater than male undercounts. Most age groups are undercounted, but the magnitude of undercounting is greater for increasingly younger ages, with more transient age groups overcounted (the oldest age groups and college-aged persons). This pattern does not appear to be site-specific but seems to be an artifact of administrative record processing and reporting lags.

Blacks and Hispanics tend to be undercounted when they are the largest minority group and overcounted when they are not. AIs have large undercounts while APIs are undercounted and overcounted by AREX site, regardless of the proportional size of APIs. The race/ethnicity ALPEs can be attributed to deficiencies in the race imputation model. Coverage rates may also have an indirect effect on the accuracy of the race/ethnicity groups because under- and overcounted age groups in the MD and CO counties may have larger proportions of particular race/ethnicity groups.

[4] The general method of imputation was applied to the Personal characteristics File (PCF) and carried over to subsequent forms of administrative records files. Further references to race imputation distinguish PCF vs. tax form methods that were applied to children less than 18 years old and acquired from the tax filer in their household.

The county-level analysis builds on the AREX-Census count results by examining the algebraic percent error (ALPE). The ALPE measure provides a different view of the county-level results because the calculation method uses group totals as bases and provides a standardized gauge for comparing differences between Top-Down and Bottom-Up, as well as between counties.

TOTAL POPULATION

- All county Bottom-Up ALPEs were smaller than Top-Down ALPEs; Bottom-Up ALPE improvements were variable: the Jefferson County Top-Down ALPE (-8.8 percent) was -3.0 percent in Bottom-Up, while the Baltimore City Top-Down ALPE (-8.8 percent) was +1.8 percent in Bottom-Up.
- The smallest total population Bottom-Up ALPEs were in Baltimore County (-1.1 percent) and El Paso County (-1.5 percent); the largest Bottom-Up ALPEs were both in CO in Jefferson (-3.0 percent) and Douglas (-3.2 percent) counties.

Bottom-up ALPEs were generally smaller due to more stringent address-matching requirements (compared to Top-Down) and Census pull households that replaced unmatched Census addresses. The overall effect provided by Bottom-Up was to increase the number of AREX households and eliminate unverified households that may place households in the wrong blocks or have unsubstantiated demographic characteristics. All of these operations, as well as estimates from the field address verification (FAV) process, may have a greater effect on population totals.

SEX

- Male and female Bottom-Up ALPEs were relatively small in all five counties and ranged from –4.0 percent to +4.2 percent.
- Both male and female proportions were undercounted in all counties and generally are unbiased, reflecting the magnitude of total county-level proportions; female undercounts were slightly worse than male undercounts and had small marginal differences in Bottom-Up.

Some women may be less active within the administrative records systems. For example, some retirement studies indicate that lifetime participation in the labor force varies by a woman's child raising and care giving experiences, health status, and race/ethnicity (Flippen and Tienda, 2000). The difference between male and female undercounts may also be attributable to delayed reporting of mortality because men and women have different survival rates at younger and older ages that vary by race and ethnicity.

AGE

- Generally, younger age groups (especially the 0-4 age group) had the largest negative ALPEs in all five counties; Bottom-Up ALPEs for the 0-4 age group ranged from –33.8 percent in Jefferson County to –23.4 percent in Baltimore City.
- Older age groups (65-74, 75-84, and 85+) tended to have positive ALPEs that increased by increasing age.

Large age-group ALPEs are likely due to the combined effect of errors in the administrative record collection process and recording lag from demographic processes. Infants are likely to have poorer coverage due to reporting lag and reporting their births. Households with five or more children, new dependents born between the beginning of tax year 1999 and the April 1, 2000 date of the Census, and separated or remarried parents who did not claim a child in their tax return are also likely to have incomplete coverage of household members. This is demonstrated by the large undercounts for the 0-4 age group.

College-age persons whose residence may have been reported at a parent's IRS tax address may actually reside on a campus in a different area. The IRS 1040 tax files also provide incomplete information for the 15 months preceding Census day, as these files are limited to 1998 tax year records. The 20-24 year age group also has large ALPE overcounts in the AREX counties. Douglas County is an extreme example where the age 20-24 Census population is about half the state and national proportions. But Colorado Springs is the home of the Air Force Academy and several universities, despite its small total population. Persons aged 65+ were generally overcounted in all five counties, which may be due to administrative records not capturing migration (to new residences and nursing homes) and mortality of older persons. Despite linkages to Medicare records, some older persons (age 65+) appear to have less reliable information in administrative records because lagged reporting may count persons alive and resident who may have died or moved. This is especially true for persons age 85+ who displayed

Bottom-Up overcounts for all five counties ranging from about two percent to 36 percent. Also, because the 85+ population is a relatively small proportion, the denominators of the ALPE calculations are likely to be small and potentially inflate ALPE measures. Generally, the 65-74 and 75-84 age group had small positive ALPEs in all five counties.

RACE

- Blacks were overcounted (Bottom-Up) in all three CO counties and Baltimore City (where Blacks are the largest minority group).
- Hispanics were overcounted (Bottom-Up) in both MD counties and undercounted in all three CO counties (where Hispanics are the largest minority group).
- American Indians had the greatest ALPEs in all five counties and were a larger proportion in the CO counties; Bottom-Up American Indian (AI) ALPEs ranged from –34.1 percent in Jefferson County to –11.3 percent in Baltimore City.
- Asian Pacific Islanders (APIs) were overcounted (Bottom-Up) in all three CO counties and Baltimore City (API proportions were similar in MD and CO counties).

Whites and Blacks are overcounted in four of the five counties (Bottom-Up results). The results further support a lack of precision in assigning White and Black races, due to deficiencies in the race imputation model and the more lenient processing of the Top-Down data. Generally, the Bottom-Up results had smaller White undercounts in four counties and smaller under- and overcounts in CO, compared to the Top-Down results. The race imputation model exhibited 'regression towards the mean' in assigning Black and White races, because aggregate population estimates were used to estimate individual race characteristics. And for AIs and APIs, the ALPE results are somewhat misleading due to the small population bases of the minority races.

The large race/ethnicity ALPEs (under- and overcounts) were probably affected by poor results from the race imputation model. The overcounting of Whites in Baltimore City and undercounting of other races is also indicative of the poor performance of race imputation. AIs had large undercounts in all counties, and despite the moderate to large proportion of imputed records, the race imputation model had little effect on the assignment of AI as a race category.[5] This undercounting may be due to a deficiency in the administrative records and their inability to accurately capture AI membership. The AI counts are unique among the race/Hispanic group measures as they reflect the smallest race category. However, unlike other administrative data sources, race information from the Indian Health Services provided the most reliable source of data, though coverage of AIs was limited to tribal and reservation populations.

The distinguishing feature between Baltimore City and the other counties is that Baltimore City has the greatest proportion of Blacks and other minorities, as well as a large proportion of older persons. The lower socioeconomic status of some Baltimore City residents may inhibit their coverage in administrative records because they may be poorly integrated with employers and federal agencies. Unemployment and not having a bank account reduces coverage in IRS 1040 and 1099 records, as well as being an active participant in Numident records. And older Blacks have been observed to have higher mortality rates than other race/ethnicity groups (Hayward and Heron, 1999). Subsequent analysis of age and race characteristics sheds some light on whether socioeconomic status and/or greater mortality contribute to the female undercount in Baltimore City.

Hispanic ALPEs had large undercounts in all three CO counties, but neither of the MD counties.

---

[5]The race model uses additional data sources for Hispanics and Asians, compared to Whites and Blacks, and uses administrative data from the Indian Health Services for AIs. Refer to Table 4.5 for imputed proportions by race and ethnicity.

Hispanics are a much larger proportion of the total population in CO (5 to 11 percent) but a smaller percent and number of the MD population (less than two percent). The results suggest that the undercount may be due to problems with race coding, the race imputation model, recent Hispanic migrants to CO (reporting lags in AREX data sources), or persons not appearing in administrative records. For example, casual labor and domestic workers may receive cash payment, provide false SSNs, and may not exist in administrative records. That is, they were captured in Census, but migration, type of employment, and AREX processing may be associated with their undercounting. APIs had large undercounts in Baltimore City only, but were overcounted in El Paso and Jefferson Counties and had smaller undercounts in the remaining counties.

One problem area with race reporting concerns the source of administrative records for persons less than 18 years old. SSA Numident records are a primary source of race/ethnicity identifiers and are generally blank for children. The Enrollment at Birth Program (EAB) does not record race/ethnicity information for new birth certificates and children lacked race identifiers in Numident. Young persons are unlikely to have any of their administrative records updated until they begin working, reach driving age, marry, or become eligible for Social Security or Medicare benefits under some catastrophic health or family incident.

An important difference between the Top-Down and Bottom-Up results was the manner in which the race imputation model treated children. In the Bottom-Up process, children were assigned the race of the primary tax filer at their address. The 1998 tax returns linked the householder and first four dependents, allowing householder race from SSA Numident to be assigned to dependents. For traditional married families, it is likely that three children plus the spouse were linked to the householder. The PCF imputation methodology was developed from a sample of adults from the Current Population Survey (CPS) and part of the improved race assignment in Bottom-Up may be due to this additional race imputation process. While a formal evaluation of the revised race imputation methodology has not been conducted, it is assumed that the more stringent Bottom-Up address requirements and household race assignment improved the accuracy of race assignment for children.

Differences between Census and AREX county-level counts can be attributed to three general causes:

- Operational factors, including, Bottom-Up/Top-Down processing, allocation from collection blocks to tabulation blocks, the race imputation model, and Field Address Verification (FAV).
- Administrative factors and their interaction with demographic events, affecting the coverage, accuracy, completeness, and timeliness of administrative data collection by federal agencies.
- Demographic factors, including mortality, fertility, and migration, and their differential impact on age groups, sex, and race/ethnicity.

4.1.2 Index of Dissimilarity Results

Summary of results: The county-level dissimilarity indices fortify the results of race and age differences and ALPEs: Bottom-Up provided better results than Top-Down and aggregate age differences exceeded race/ethnicity differences.

The index of dissimilarity provides a single measure of correspondence between two different distributions and summarizes race/ethnicity and age group differences between AREX and Census. A greater index indicates one or more race/ethnicity or age categories differs between AREX and Census within the county, but does not distinguish which particular category is different. The indices are sensitive to the number of groups and group ranges used. There are more age groups than race/Hispanic groups so the age dissimilarity index is slightly larger. This section describes county-level results, while subsequent comparisons between county and other geographies use identical group definitions to facilitate comparisons across geographies.

- Bottom-up had smaller indices, compared to Top-Down, and were significantly smaller in Baltimore City.
- The Bottom-Up age dissimilarity index exceeded the race dissimilarity index in all counties except El Paso County, which had the smallest Bottom-Up index among the CO counties.
- The Bottom-Up age and race dissimilarity indices were generally greater in all CO counties, compared to the MD counties.

These results mirror what was found in the individual age, race, and Hispanic comparisons. The reduction in the race dissimilarity index from Top-Down to Bottom-Up is significant, considering that these measures reflect the largest and smallest indices of all calculated county indices. In all comparisons except the Jefferson County age index, the Bottom-Up method provided more accurate results than Top-Down. The AREX race counts approximated Census results in Douglas County, while age was better measured in El Paso and Jefferson Counties. The Bottom-Up results support that the age dissimilarity index is somewhat constant across the counties, suggesting that Bottom-Up was neutral in its treatment of age groups. However, the age dissimilarity index is an aggregate measure and the age-group components may offset each other, because one component of the index might be very large and dominate the summed value. The treatment of race across counties was varied, and the large reduction in the race dissimilarity index in Baltimore City merits further investigation.

4.2 State Legislative District Comparisons

Summary of results: The comparison of state legislative districts and Census results emphasizes Bottom-Up household counts, and are compared to Census 2000 results that include persons in households. The state legislative districts show remarkable heterogeneity given the size of each district. The number of overcounted districts exceeded undercounted districts in both sites, though the undercounted districts had small magnitudes. The chief difference between county and state district results is the exclusion of persons under 18 years old.

The comparison between AREX population estimates of state legislative district Bottom-Up counts and Census 2000 household results focuses on the age 18+ populations of the districts. This simplified analysis will compare AREX-Census total population differences and ALPEs. The legislative districts are composed of census blocks and can flow across county boundaries. Consequently, the comparisons focus on districts that are wholly contained or large parts of districts that lie within the MD and CO AREX sites. The comparison is somewhat biased because it pits AREX households against all Census persons and excludes GQs. It is assumed that efforts beyond the current administrative records methods in this evaluation will be required to accurately count persons in GQs. Consequently, the AREX household population is used as an estimator of district-level total population counts.

The AREX-Census counts, differences, and ALPEs by legislative district are shown in Table 4.2. Disaggregating the county counts reveals the heterogeneity of the district-level counts, as well as under- and overcounting by AREX. All of the county-level AREX counts were less than Census, but nearly all of the district-level AREX counts exceeded Census results. This is due to the exclusion of persons under aged 18. The range of district-level ALPEs is wider than the county-level ALPEs. The results for the CO site were similar to the MD results, as each site had several districts with ALPEs exceeding county results.

One criterion for redistricting is equal size, where the total population of each legislative district is within five percent of a pre-specified value. Because some of the districts are incomplete and reflect uncounted persons, it is not possible to test the extent that the districts met this criterion. However, 80 percent of districts had AREX counts within five percent of the Census values for these household counts.

Table 4.2: Voting Age Persons (Aged 18+) by State Legislative Voting Districts

Maryland Districts	AREX	Census	Difference	ALPE	Colorado Districts	AREX	Census	Difference	ALPE
5B	25612	24039	1573	6.5%	14	46715	45490	1225	2.7%
6	86517	84347	2170	2.6%	15	47993	46800	1193	2.5%
7	69103	69478	-375	-0.5%	16	49796	48705	1091	2.2%
8	88706	86755	1951	2.2%	17	40948	39811	1137	2.9%
10	87760	85051	2709	3.2%	18	51294	50059	1235	2.5%
11	85791	81956	3835	4.7%	19	46848	43703	3145	7.2%
12A	46664	45024	1640	3.6%	20	44413	41225	3188	7.7%
31	7394	7196	198	2.8%	21	45176	43260	1916	4.4%
40	75375	71076	4299	6.0%	22	47870	46754	1116	2.4%
41	82248	77279	4969	6.4%	23	47811	47902	-91	-0.2%
42	80348	80539	-191	-0.2%	24	50252	50433	-181	-0.4%
43	78623	74480	4143	5.6%	25	49300	47623	1677	3.5%
44	82407	78515	3892	5.0%	26	51167	50319	848	1.7%
45	79910	75171	4739	6.3%	27	50308	48495	1813	3.7%
46	84226	81431	2795	3.4%	28	47969	45206	2763	6.1%
TOTAL	1060684	1022337	38347	3.8%	29	49367	47671	1696	3.6%
					33	1280	1244	36	2.9%
					38	209	191	18	9.4%
					43	44542	43917	625	1.4%
					44	45536	44226	1310	3.0%
					45	33210	31680	1530	4.8%
					TOTAL	892004	888306	3698	0.4%

4.3 Tract Comparisons

Summary of results: The comparison of tract and Census results focuses on Bottom-Up counts. The tract-level ALPE results indicated a good correspondence between AREX and Census total population counts (72 percent of tracts met the five percent criterion and 99 percent met the 25 percent criterion), though some tracts had moderate and large ALPE undercounts. ALPE results for sex and age were similar for tract and county analyses, with smaller under- and over-counts associated with larger proportions of accurately counted tracts. Baltimore City had the worst results for total and demographic ALPE measures but the most accurate results for Blacks. However, Baltimore City also had the largest proportion of census pull records and smallest proportion of imputed Black race codes. For the race/Hispanic minority groups, the relative size of the minority population was associated with how well AREX simulated Census results, because small minority proportions tended to have more tracts with moderate or large ALPE overcounts. Because of methodological differences between the tract and block analyses, the general analyses of block-level ALPE distributions are not discussed.

The AREX-Census tract comparisons emphasize ALPE Bottom-Up distributions and use the same population demographics described in the county-level analyses. Because the Bottom-Up results were found to be more favorable in the county-level results, Top-Down results are not presented. The tract comparisons provide a unique set of problems if processing errors accumulate from the race imputation model, FAV estimation, and allocation from collection to tabulation block processes. For example, if a contiguous group of blocks have under- or overcounts and these blocks are aggregated into tracts, then the resultant tract could have a significant under- or overcount. Mean tract errors become inflated because the extreme values of some tracts may behave like outliers and inflate site-level descriptive statistics. All tract value differences that exceed the 95th percentile have been topcoded or trimmed to equal the value of the 95th percentile. While topcoding can alter the magnitude of AREX-Census ALPEs, cumulative processing errors remain in the data and may seem to conflict with the county-level results. Comparing tract and county ALPE results provides information on the accuracy of tract-level characteristics relative to counties. But the main problem with this type of comparison is that the ALPE denominator potentially inflates tract-level ALPEs for small population subgroups and especially minorities. See Appendices 3 and 4 for more discussion on tract and block incongruities.

TOTAL POPULATION

- More than 70 percent of tracts had AREX total population counts within +/- five percent of Census results (five percent criterion), and more than 95 percent of tracts had counts within 25 percent of Census results (25 percent criterion) in four of five counties; Baltimore City had the least accurate results with 50 percent of tracts exceeding +/- five percent of Census results.
- A larger proportion of tracts had moderate and large undercounts (less than +/– five percent) compared to overcounts (results not shown).

Though the tract-level ALPEs for the total population resemble county-level results, the distributions indicate more Baltimore City tracts were overcounted. It's unclear whether these overcounts are related to persons who were actually uncounted in Census, or more likely, flaws in AREX processing. Households may have been added through the Census pull households that replaced unmatched addresses that existed in other tracts or addresses.

SEX

- A larger proportion of CO tracts had male ALPEs within the five percent criterion, compared to females; in both MD counties, and especially Baltimore City, female ALPE results were more accurate than male results.
- Baltimore City had the least accurate correspondence between AREX and Census at both five percent and 25 percent ALPE criteria.

The sex ALPE results reflect the tract-level total population counts. The most important issue is whether AREX counted males or females more accurately. There are several possible explanations for why tract-level accuracy varies by sex. Low-income women in urban areas may have weaker links to the economic institutions of larger society and poorer coverage in administrative data sources. Under coverage may also be due to working men and women who did not contribute to Social Security and tax rolls. Older women may have larger undercounts or smaller overcounts than older men because older women are more likely to outlive their husbands and exist in Census and AREX data. But older women may migrate near their children or other relatives, or enter nursing homes. That is, older women may be counted in AREX but not at the same Census address, and potentially offset expected female undercounts.

AGE

- The 45-54 and 55-64 age groups had the most accurate AREX counts; about 70 to 80 percent of tracts counted persons aged 45-64 within the five percent criterion; Baltimore City was somewhat worse with less than 60 percent of tracts within the five percent criterion.
- Age groups with the smallest proportion of tracts within the five percent criterion included ages 0-4, 20-24, and 85+; age groups 0-4 and 85+ had the smallest proportion of accurate tracts at the 25 percent criterion; the results were similar for the 20-24 year age group in four of five counties (Douglas County had less than 30 percent of tracts within the 25 percent criterion).
- Generally, about 90 percent of tracts were accurately counted at the 25 percent criterion for ages 25-74.
- The distribution of ALPEs covering ages 5-64 indicate that under- and overcounts are affected by moderate differences (five to 25 percent) between AREX and Census counts, rather than large errors.
- For older age groups, there was a strong association between increasing age and decreasing accuracy of AREX results at both the five percent and 25 percent criteria.
- Despite the small proportion of tracts within the five percent criterion for the 65-74 and 75-84 age groups, about 90 percent of tracts were counted accurately at the 25 percent criterion.
- The 85+ year age group had the largest proportion of overcounts and largest proportion of large overcounts (exceeding the 25 percent criterion).

The tract-level age interval undercounts corresponded with the county-level Bottom-Up results. Age groups 0-4, 20-24, and 75+ were measured less accurately within tracts, and the most extreme age groups, 0-4 and 85+ had the smallest proportion of tracts within the five percent or

25 percent criteria. The substantive implications of the tract-level ALPEs support the county-level results: younger age groups tend to be undercounted because they have less exposure to administrative record-keeping agencies and the limitations of the IRS 1040 tax records, while reporting lag affects the accuracy of tract-level results of older age groups because of lagged reporting of mortality and migration. The linkage between 20-24 year olds and their parents' tax returns, and the generally higher mobility rates for young adults reflects the county-level ALPE results, though disaggregation from county to tracts confounds the relationship.

RACE/ETHNICITY

- Whites had the largest proportion of accurate results in four of five counties, ranging from about 45 to 90 percent of tracts meeting the five percent criterion; about 25 percent of tracts in Baltimore City had accurate counts for Whites at the five percent criterion.
- Blacks were counted more accurately in the MD counties and Hispanics were counted more accurately in the CO counties, at both five percent and 25 percent criteria; Hispanics in the MD counties had a significant proportion of tracts with large overcounts (exceeding the 25 percent criterion), while Blacks in the CO counties also had a significant proportion of tracts with large overcounts.
- Generally, race groups with the smallest population proportions had the smallest proportion of tracts within the five percent and 25 percent criteria.
- Baltimore City, with a Black majority, was most accurate in counting Blacks, compared to the other race groups.
- Counties with small proportions of a particular race tended to have more tracts with moderate or large overcounts for that race (for example, Hispanics in Baltimore County); this caused some county-level results that showed undercounts to appear as tract-level overcounts (for example, Hispanics in Douglas County).
- AIs had the least accurate results of the race groups and the greatest proportions of large under- and overcounts (exceeding –25 percent; see appendix).

In general, the direction of tract-level mean ALPEs corresponded with county-level results. Race categories with small proportions, especially AIs and APIs, were more likely to differ between geographies and have different magnitudes when they did correspond. This was due to larger errors in more tracts. Counties with fewer tracts also had greater errors and less correspondence with county-level results, especially Douglas County.

The accuracy of tract-level race counts is affected indirectly through the age composition of tracts, and directly through the race imputation model. As was found in the review of the age category results, the youngest, oldest, and early adult age categories had the least accurate results. Because the race categories have very different birth, death, and immigration rates, the age category errors are likely to impact the race groups in different ways. For example, if Black and Hispanic fertility rates are greater, compared to other race groups, then Blacks and Hispanics may have greater net undercounts for this age group. Similarly, Black mortality at older ages is higher than other race categories and may produce a larger net Black overcount.

The race imputation model has been found to produce good estimates of national race proportions, but poor estimates for small areas. Some of the tract-level errors may be attributable to the poor performance of the race model. Multivariate analyses that distinguish the source of the assigned race are performed in this report and attempt to decompose the influences of age composition and race imputation model on race category results. However, residential segregation is likely to produce neighborhood clusters of errors, associated with the race-mix of neighborhoods and the number of contiguous blocks and tracts. Spatial analyses provide further elaboration of the distributional characteristics of tract and block-level results and are presented in a later section of this report.

AIs have a separate source of administrative data, though AIs in urban areas not integrated with the Indian Health Services are likely to be less accurate. Despite the moderate to large proportion of imputed AI race results, the large undercounts suggest that the race imputation model provides an inaccurate assignment of AI race status, compared to the other race groups. However, AIs were generally the smallest race proportion and ALPE calculations with small denominators produce larger ALPEs. Whites in Baltimore City had large overcounts that were not reflected in county-level results. One additional problem with the race imputation model is its inability to distinguish Whites and Blacks. This issue is more problematic in Baltimore City with its Black majority population.

4.4 Block-level ALPEs

The block-level ALPE results describe the accuracy of counts at the smallest geographic level and relative to counties and tracts. The main problem with this type of comparison is the ALPE denominator potentially inflates block-level ALPEs for small population subgroups and especially minorities. This inflation is likely to be greater than found in the tract-county comparisons. A second issue affecting comparisons is the exclusion of blocks where census did not identify persons with a particular attribute (zero blocks). Tract and block ALPEs include blocks with zero counts because these blocks were collapsed into larger geographies. However, the block-level ALPEs use the reduced sample of blocks and the results may be quite different when comparing the ALPEs at various geographies.

- AREX was more accurate in estimating tracts than blocks in all counties; from about 25 to 40 percent of blocks were within the five percent criterion, and about 85 percent were within the 25 percent criteria in the five counties; Douglas County had the best results at the five percent criterion and Baltimore County was best at the 25 percent criterion.
- In the MD counties, slightly more blocks had moderate or large overcounts (ALPEs exceeding five percent), compared to the CO counties where more blocks had moderate undercounts (minus five to –24 percent).

The AREX counts were less accurate at the block-level. Total population proportions are likely to be less accurate at smaller areas due to incorrect assignment of households at tracts and blocks that average out for county-level counts. This is demonstrated by the greater number of moderate and large ALPEs and indicates how smaller denominators and AREX processing flaws influenced the results. Though zero blocks were excluded and fewer blocks met the five percent criterion, a surprisingly large proportion of blocks met the 25 percent criterion in all five counties. Block-level results for sex, age, and race can be found in Appendix 5.

4.5 AREX Processing and Operational Issues

Race assignment can be decomposed into three major categories:

- Most frequent report from source administrative files.
- Imputed from PCF probability estimates and assigned to adults.
- Imputed from householder's race and assigned to children under 18 years old.

Table 4.3 provides a summary of race imputation and Census pull proportions; more detailed tables can be found in Appendix 2. The imputed race assignments may increase AREX-Census differences while Census pull improves the apparent accuracy of AREX. The distribution of imputed and Census pull cases fall into several distinct patterns, though later analyses identify how the race assignment process affected ALPE results:

- Race imputation was greater in the CO counties, especially for Whites and Blacks.
- Both MD counties had similar imputation rates, though the rate of Census pull was much greater for Baltimore City.
- Jefferson and Douglas Counties had the greatest imputation rates for total population and most of the race categories, while Douglas County and Baltimore City had the greatest Census pull rates.

Table 4.5: Summary of Race and Hispanic Origin Imputation Rates by County[1]

Race/Hispanic Category	Baltimore County	Baltimore City	Douglas County	El Paso County	Jefferson County
All persons	12.5%	9.8%	17.1%	16.9%	17.4%
White	12.0%	11.0%	16.8%	16.9%	15.6%
Black	11.7%	8.9%	26.7%	15.3%	31.6%
AI	28.8%	24.2%	26.3%	20.2%	21.1%
API	28.7%	20.7%	28.2%	27.4%	31.2%
Race Unknown	0.2%	0.1%	0.2%	0.8%	0.6%
Hispanic	92.5%	82.6%	84.6%	85.3%	88.2%
Census Pull	6.3%	15.3%	13.5%	9.3%	7.4%

[1](Imputed PCF + householder-assigned records to children) / total AREX records

4.6 Geospatial Distributions of AREX-Census Bottom-up Differences

Figures 4.6.1a and 4.6.1b show the ALPEs for the total population of each AREX site. The intervals used for thematic mapping use a natural-break algorithm (Jenks and Caspall, 1971). Selected age and race ALPEs are shown for persons aged 0-4, 85+, Blacks, and Hispanics in Figures 4.6.2a-4.6.3b. Indices of dissimilarity in Figures 4.6.4a-4.6.5b provide a general perspective on the aggregate age and race characteristics of the AREX tracts. These final maps have different measurement scales and the intervals and color scheme differ from the previous maps.

Mapped ALPE results for total population

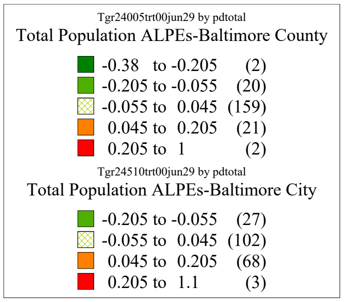
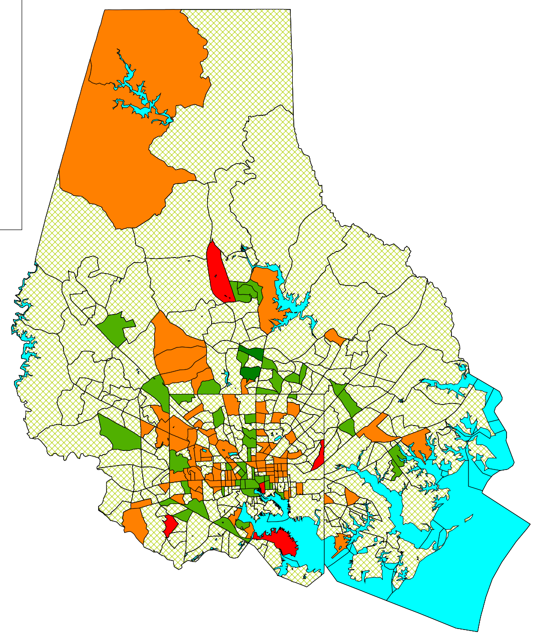


Figure 4.6.1a: AREX-Census ALPEs for the Total Population: MD Tracts



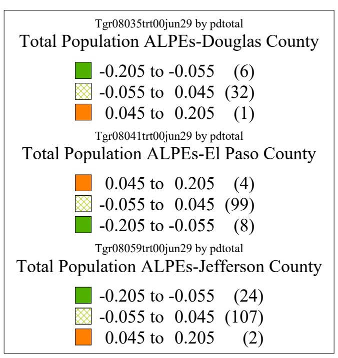
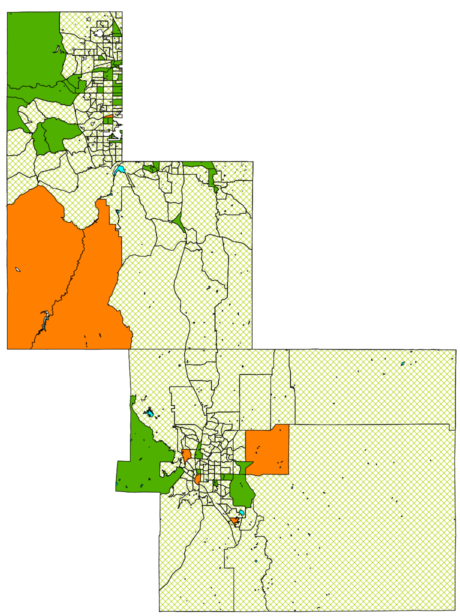


Figure 4.6.1b: AREX-Census ALPEs for the Total Population: CO Tracts



- ALPE results were better in suburban than urban and rural tracts in both AREX sites.
- Under- and overcounted tracts tended to cluster, suggesting adjacent tracts had similar population characteristics.

Mapped ALPE results for selected age characteristics

Figure 4.6.2a: AREX-Census ALPEs for Persons Aged 0-4: CO Tracts

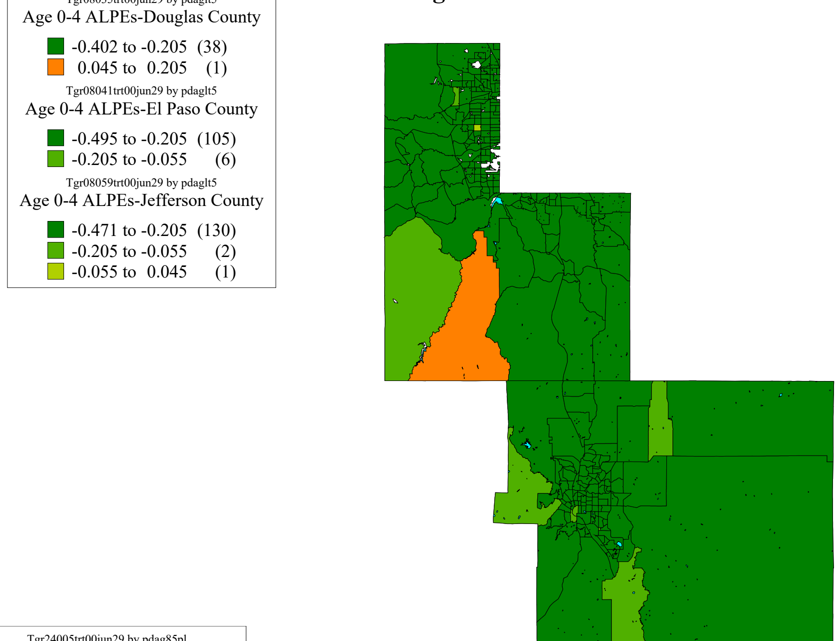
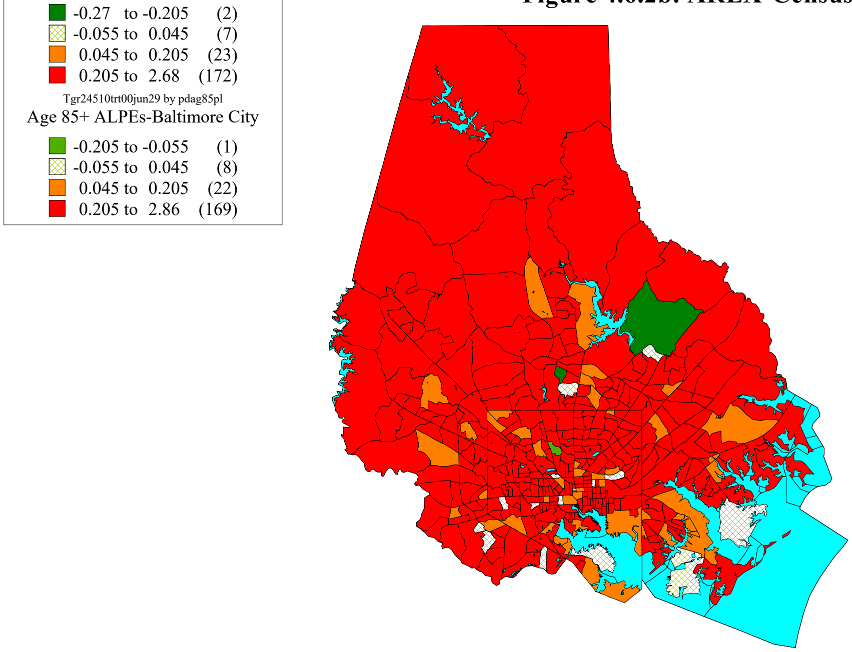


Figure 4.6.2b: AREX-Census ALPEs for Persons Aged 85+: MD Tracts



- For the age 0-4 undercounts and age 85+ overcounts, the map display reinforces that there is no tract-level heterogeneity because more than 90 percent are under- or overcounted.

The large number of under- and overcounts suggest serious deficiencies in the AREX data and/or processing. That is, federal agencies are unable to quickly incorporate demographic events such as birth and death or the administrative coverage periods do not provide a good estimate of Census day enumeration for these age groups.

Mapped ALPE results for selected race characteristics

Figure 4.6.3a: AREX-Census ALPEs for Blacks: MD Tracts

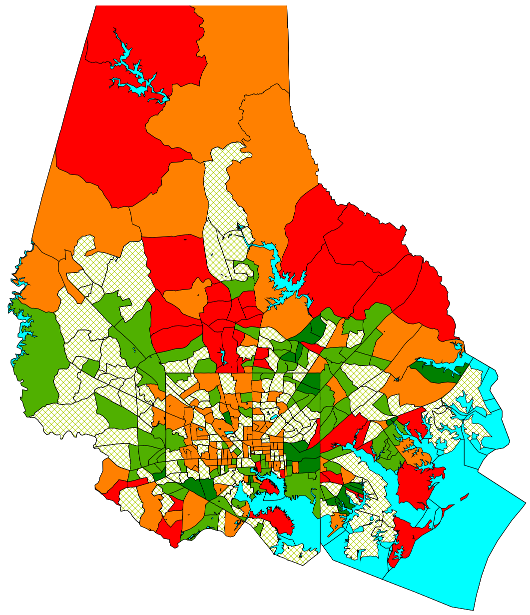
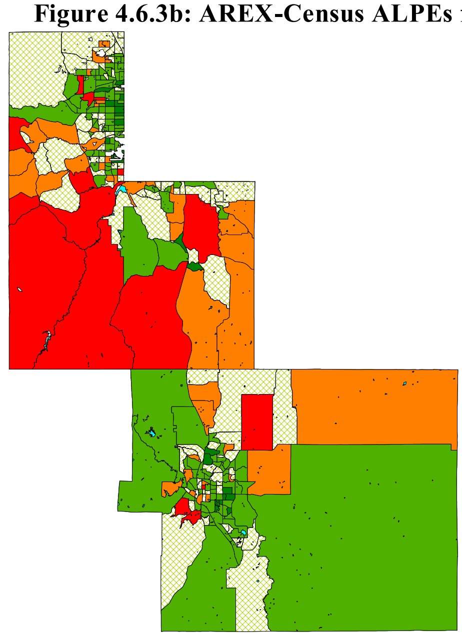


Figure 4.6.3b: AREX-Census ALPEs for Hispanics: CO Tracts



- Large Black and Hispanic ALPEs occur in urban and rural tracts in both MD and CO counties, with large overcounts more frequent in rural areas and large undercounts in urban areas.
- Moderate and large under- and overcounts were similar in both MD and CO counties.

Black overcounts in the MD counties are probably due to errors in AREX processing, especially the race imputation model, that incorrectly assigned Blacks to tracts. These larger overcounts tend to be in predominantly rural, White tracts in northern Baltimore County. Moderate overcounts and undercounts are concentrated in Baltimore City and contiguous tracts surrounding Baltimore City. However, there are some tracts with large undercounts, including small clusters within Baltimore City and around the Towson area in central Baltimore County. The spatial distribution of under- and overcounts in the CO counties is confounded by the large rural tracts that appear to weight the graphic presentation. For Hispanics in CO, there are considerably more tracts with large and moderate undercounts, some of which are large clusters and others that are isolated.

The key finding from the Black spatial distributions for the MD counties is that the overcounts are probably in error because they appear in largely rural, White areas, while large undercounts are not randomly distributed and indicate other underlying causes. The spatial-race under- and overcount patterns could be due to historic settlement and migration patterns that facilitated greater racial integration in CO, or differences in the age structures of the two AREX sites. In both sites, age, cohort, and related factors are potential contributors to spatial variations.

There also appears to be a relationship between urban and rural tracts and under- and overcounts. The large overcounts in both AREX sites appear in predominantly rural tracts and provide additional support for deficiencies in the race imputation model. Further investigation would provide more details about the impact of resident cohorts, settlement/migration patterns, and age structure on the AREX-Census differences.

Mapped dissimilarity indices

Figure 4.6.4a: AREX-Census Index of Dissimilarity for Age: MD Tracts

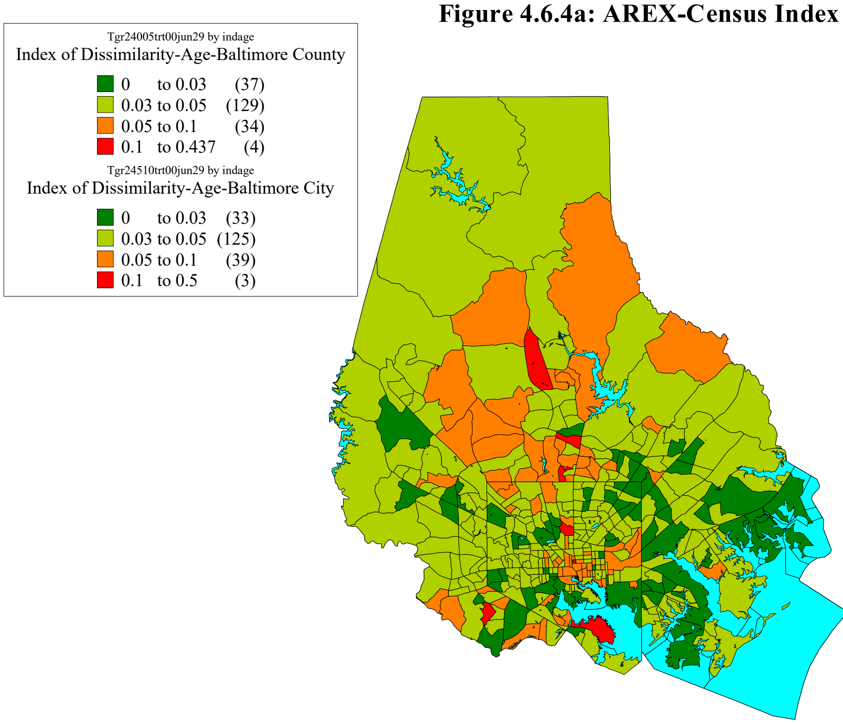


Figure 4.6.4b: AREX-Census Index of Dissimilarity for Age: CO Tracts

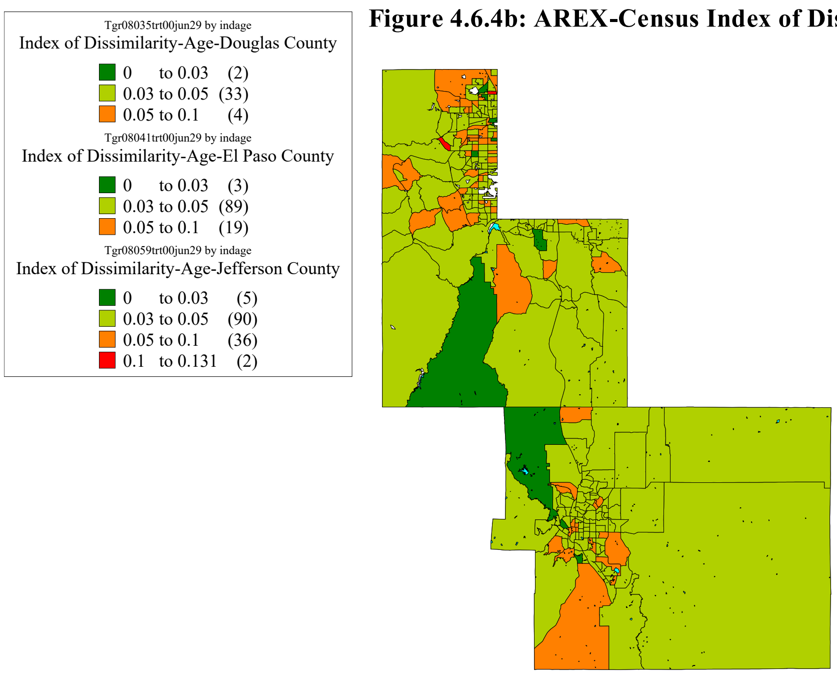


Figure 4.6.5a: AREX-Census Index of Dissimilarity for Race: MD Tracts

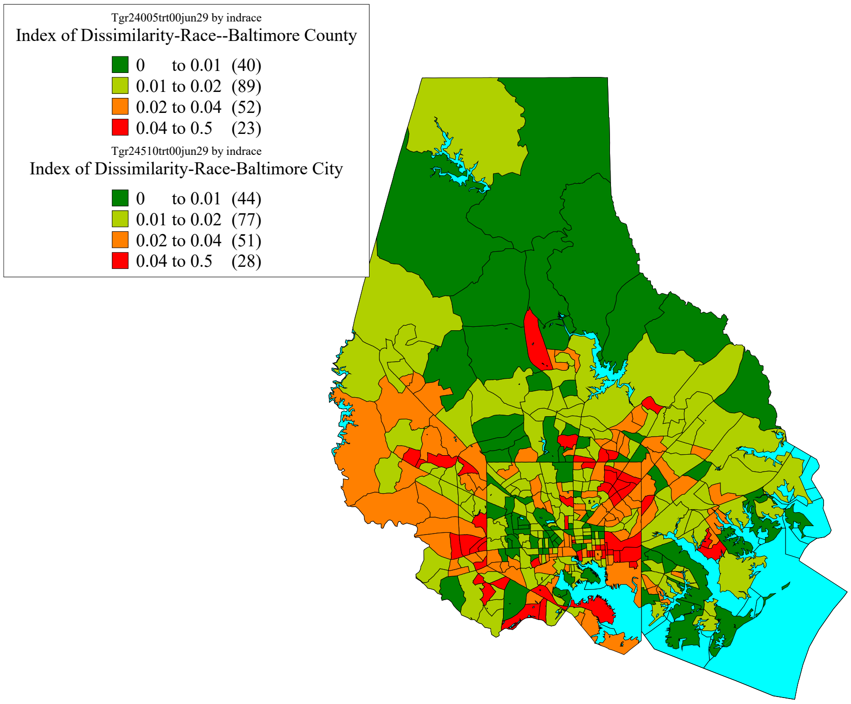
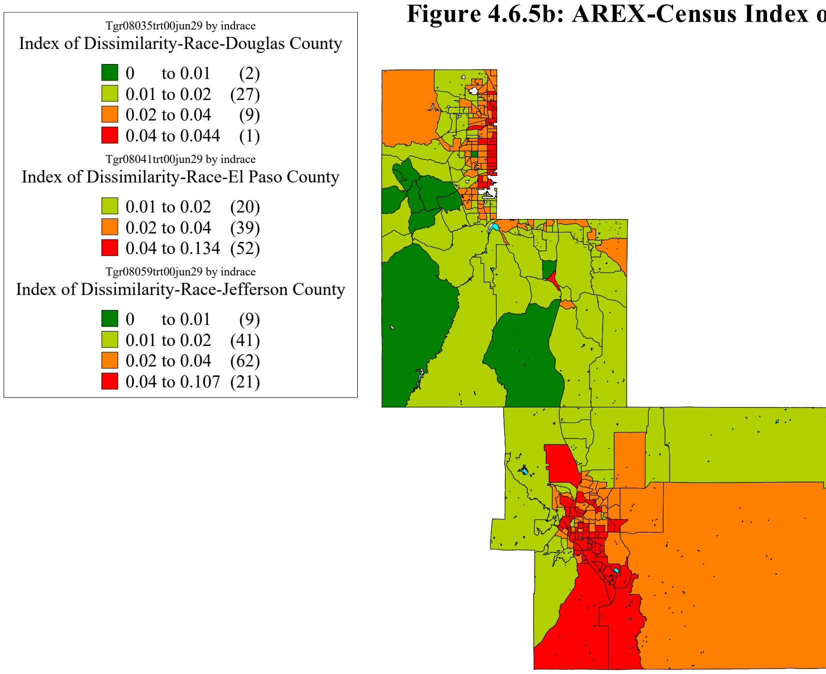


Figure 4.6.5b: AREX-Census Index of Dissimilarity for Race: MD Tracts



- Tracts with greater age dissimilarity indices are more clustered in MD than CO counties.
- Race/Hispanic dissimilarity indices are greater in urban and adjacent areas of the MD and CO counties and appear as contiguous tract clusters; in CO, greater race dissimilarity indices occur in more urbanized tracts.
- Most tracts have either high race or age dissimilarity and not both.

The spatial distribution of tract-level dissimilarity indices for age may be related to neighborhood cohort characteristics and family formation. Cohort characteristics reflect predominantly Black or White residents who assume households at early ages and remain there through retirement. Family formation reflects children who were undercounted and are also likely to be associated with specific cohorts and neighborhoods.

Despite similar population sizes in Baltimore City and County, 90 percent of the high race/Hispanic tract indices in MD are in Baltimore City. This is primarily due to the large minority population in Baltimore City. However, AREX-Census differences in Baltimore City are localized in four tract clusters. The other notable issue in Baltimore City is that some of the low dissimilarity tract clusters are in predominantly Black neighborhoods. That is, not all minorities and Blacks have been poorly represented by AREX counts. The key issue is what non-race/Hispanic attributes are contributing to the greater dissimilarity between AREX and Census for minorities. Tracts with large race/Hispanic or age dissimilarity indices could be associated with the price and availability of housing stock, the population demographics in those clusters, as well as the non-resident characteristics of those clusters, including schools, highways, and commercial districts. Racially segregated neighborhoods may also be contributing to these results in Baltimore City.

4.7 Multivariate Analysis

Summary of results: The model results confirm some of the key findings from the univariate and bivariate analyses. Among the mobility variables, both vacancy rate and rental rate were associated with under and overcounts. Rental rate had a greater impact on undercounts and vacancy rates impacted overcounts in both AREX sites. There was also a general tendency for greater imputation rates to be associated with overcounts. While the imputation rates did not affect total AREX counts, they indicated a characteristic of those blocks may be linked with AREX overcounts. Similarly, presence of multi-race and some other race reports was strongly associated with undercounts and overcounts, as indicators of some unobserved characteristic of those blocks. White, Black, and Hispanic presence and proportions had variable associations with under- and overcounts. As observed in the bivariate analyses, large proportions of persons under age 5 and 20-24 were associated with undercounts in both sites. And in CO, large proportions of persons age 65+ were associated with overcounts.

4.7.1 Categorical Model Results

The primary goals of the categorical regression models are to identify the key predictors associated with under- and overcounts and account for population composition differences between counties and AREX sites. The extensive univariate and limited bivariate analyses are confounded by the uncontrolled characteristics of blocks, tracts, and counties. That is, demographic, ecological, and socioeconomic characteristics. The multivariate models remove this confounding so that comparisons can be made between predictor variables and counties.

The model results identify the key predictors of the selected age, race/ethnicity, sex and total population ALPEs and assume the Census results to be the 'truth' about the AREX population. The regression models have been structured to answer the question: *What block characteristics are most important for understanding differences between AREX and Census results, using the Census results as a standard?* A secondary goal of the multivariate models addresses how AREX counts can be improved to more accurately depict the Census population. This can be accomplished in two ways. First, is to understand the operational and administrative deficiencies that affect the AREX counts. The operational deficiencies can be addressed internally by PRED through enhanced processing methods. Addressing administrative deficiencies is more problematic because it requires the cooperation of other federal agencies whose requirements may be at odds with the Census Bureau's desired changes. Another alternative is to use the model results and develop correction factors for the AREX counts, based on the Census results. This would be most useful for intercensal estimates that employ administrative records, but fraught with the usual problems of estimating small areas with statistical methods (Smith and Shahidullah, 1993).

Comparison of moderate and large under- and overcounts to 'best' results

The models use categorical regression methods to compare how the predictors contribute to moderate and large deviations from the 'best' results or reference group. The five categories based on the interquartile range have the ALPE ranges shown in Table 4.7.1:

Table 4.7.1: Under- and Overcount Groups for Total Population ALPEs

Group and Relative Size	ALPE Range	
	MD	CO
G1: Large undercount	< -14.4%	< -16.7%
G2: Moderate undercount	-14.4% to -2.3%	-16.7% to -4.2%
G3: Reference range	-2.3% to +5.5%	-4.2% to +2.0%
G4: Moderate overcount	+5.5% to +19.8%	+2.0% to +16.2%
G5: Large overcount	> 19.8%	> 16.2%

The model results are summarized in Tables 4.7.2 and 4.7.3 (parameter estimates in Appendix 5, Tables A5.1-A5.2). The models compare blocks with moderate and large under- and overcounts to a reference group of blocks whose AREX counts are closest to Census results (the 'best' or reference group). There are also reference demographic characteristics that were assumed to have the smallest ALPE results. The demographic reference group includes blocks with low mobility rates (low vacancy, rental, nonrelatives), mean imputation and Census pull rates, suburban or moderate population density, moderate White population proportions, no mention of Blacks or Hispanics, and a large proportion of persons aged 45-64.

While the interpretation of results appears confusing, the focus in this evaluation is a general understanding of variables relationships affecting under- and overcounts. For example, a low vacancy rate in MD (less than the median) is associated with a large undercount, relative to the 'best' AREX results. And there is a clear trend between vacancy rate and under- and overcounts. Low vacancy rate is associated with moderate undercounts, though the strength of the association is less. And higher vacancy rates contribute to increasing ALPE overcounts. Several sets of findings can be derived from the model results that answer the questions:

What effect did a particular variable have on moderate and large under- and overcounts, relative to the reference group?

What trends and relationships exist for a particular variable across the under- and overcount groups?

How did the variable relationships differ between the MD and CO sites, as well as by counties within sites?

Tables 4.7.2 and 4.7.3 provide a summary of the results to help answer these questions.

Table 4.7.2: Summary of Categorical Model Results-MD

Large undercounts	Moderate undercounts
Low vacancy rate	High rental rate
High rental rate	Large proportion of non-relatives in household
Small proportion of imputed race	Small proportion of imputed race-tax method
Small proportion of imputed race-tax method	Large proportion of imputed ethnicity
Large proportion of imputed ethnicity	Large proportion of Census pull cases
No multi-race reports	Small proportion of low density blocks
High population density	Black presence
Neighborhoods 3 and 4	Large proportion of persons under age 44
Large proportion of persons under age 5 and 20-24	Baltimore County
Small proportion of persons age 65+	
Baltimore City	

Moderate overcounts	Large overcounts
High vacancy rate	High vacancy rate
Large proportion of imputed race	High rental rate
Large proportion of imputed ethnicity	Small proportion of non-relatives in household
Large proportion of Census pull cases	Large proportion of imputed race
No multi-race, some other race reports	Large proportion of imputed ethnicity
Small proportion of low density blocks	No multi-race, some other race reports
Neighborhood 3	Low and high population density, Q1, Q2, Q5
Small proportion of persons under age 20	Small and large proportion of Whites, Q1, Q2, Q5
	Small proportion of persons under age 20

Characteristics Impacting Under- and Overcounts
Vacancy rate
Imputed race-tax method
Imputed ethnicity-large for all models
Census pull affected moderate under- and overcounts only
Large proportion of Whites (Q5)
Age < 5 and 20-24

Table 4.7.3: Summary of Categorical Model Results-CO

Large undercounts	Moderate undercounts
Smaller undercounts in El Paso than Jefferson County	Less in El Paso, more in Douglas than Jefferson County
High rental rate	High rental rate
Large proportion of non-relatives in household	Large proportion of non-relatives in household
Small proportion of imputed race-pcf	Large proportion of imputed race-PCF method
Large proportion of imputed ethnicity	Large proportion of imputed ethnicity
No multi-race reports	Large proportion of Census pull cases
High population density	No multi-race, some other race reports
Large proportion of persons under age 24	High, not low population density
Large proportion of persons age 65+	Not neighborhood 2
	Large proportion of persons under age 24
	Large proportion of persons age 65+

Moderate overcounts	Large overcounts
High vacancy rate	High vacancy rate
Large proportion of non-relatives in household	High rental rate
Large proportion of imputed race (pcf and tax)	Large proportion of imputed race (pcf and tax0
Large proportion of imputed ethnicity	Large proportion of imputed ethnicity
No multi-race, some other race reports	No multi-race, some other race reports
Not high population density	Low population density
Not neighborhood 4	Small and large proportion of Whites-Q1, Q2, Q5
Small proportion of persons under age 5, 25-44	No Hispanic presence
Large proportion of persons age 20-24, 65+	Small proportion of persons under age 19
	Large proportion of persons age 20-24, 65+

Characteristics Impacting Under- and Overcounts
Vacancy rate
Rental rate
Nonrelatives
Imputed race-tax and pcf
Imputed ethnicity-large for all models
Neighborhood 2- undercounts only
No Hispanics have large under- and overcounts
Age <5, 5-19, 65+

Piecewise regression models on ALPE values

The analyses in this section treat each subset of cases in an interquartile group as a separate multivariate regression model. The purpose is to minimize the prediction error in the models so that regression residuals can be calculated and presented in thematic maps. The same predictor variables used in the categorical models are used for the piecewise ALPE models. The dependent variable in each model is the actual block-level ALPE. In addition to total ALPE, Black, Hispanic, and age ALPE models are estimated. The model results (see Appendix 5, Tables A5.3-A5.8) describe how well total Census counts were estimated by AREX, but also show how the most critical race and age categories were affected by mobility, imputation, and demographic variations in the AREX sites.

Total Population ALPEs in MD and CO sites:

- The vacancy, rental, multi-race, and some other race variables had smaller effects on the actual ALPE measures than indicated in the categorical models.
- All of the imputation and Census pull measures were associated with large undercounts; in MD, the imputation variables were also associated with overcounts.
- Proportion of Whites was a strong predictor in CO for all ALPE ranges.
- Age variables were more important in CO than MD, with large proportions of persons in the youngest and oldest age groups were associated large undercounts.

Age 0-4 ALPEs in CO counties:

- The tax imputation method and Census pull variables were associated with moderate undercounts and the reference or 'best' ALPE group.
- Lack of multi-race and blocks without Black residents were associated with large overcounts.
- Small proportions of age 0-4 persons were associated with overcounts.
- Large proportions of age 65+ persons were associated with under- and overcounts.

Age 65+ ALPEs in MD counties:

- High vacancy rates were associated with large undercounts.
- Both tax imputation methods, Census pull, and multi-race presence were associated with large undercounts.
- Low population density (Q1) was associated with moderate and large overcounts.
- Large proportions of persons age 5-19 were associated with large undercounts.

Black ALPEs in MD counties:

- Presence of non-relatives was strongly associated with moderate overcounts.
- PCF imputation and Census pull variables were associated with large undercounts.
- Race variables were only associated with undercounts.
- Low population density (Q1) was associated with overcounts, especially moderate ones.
- Large proportions of persons aged 0-4 were associated with large undercounts.

Hispanic ALPEs in CO counties:

- Presence of non-relatives and larger proportion of rental units were associated with large undercounts.
- Both tax and PCF methods of imputation were associated with under and overcounts, especially moderate overcounts.
- Presence of multi-race and some other race reports were associated with large undercounts.
- Presence of Blacks was associated with large and moderate undercounts.
- Both Douglas and El Paso Counties had larger overcounts, compared to Jefferson County.
- Only the age 5-19 group was important and predicted large undercounts.

4.7.2 Analysis of Regression Residuals

The spatial maps in this section identify zero blocks, where population values from Census 2000 are zero and ALPEs were not calculated, and small vs. moderate/large regression residuals. These block-level thematic maps also show tract boundaries and attempt to explain why there were differences between the bivariate block and tract-level results. Moderate and large under- and overcounts are not distinguished in order to simplify the presentation.

Total Population ALPE Residuals

Figure 4.7.1a: Total Population ALPE Residuals-MD

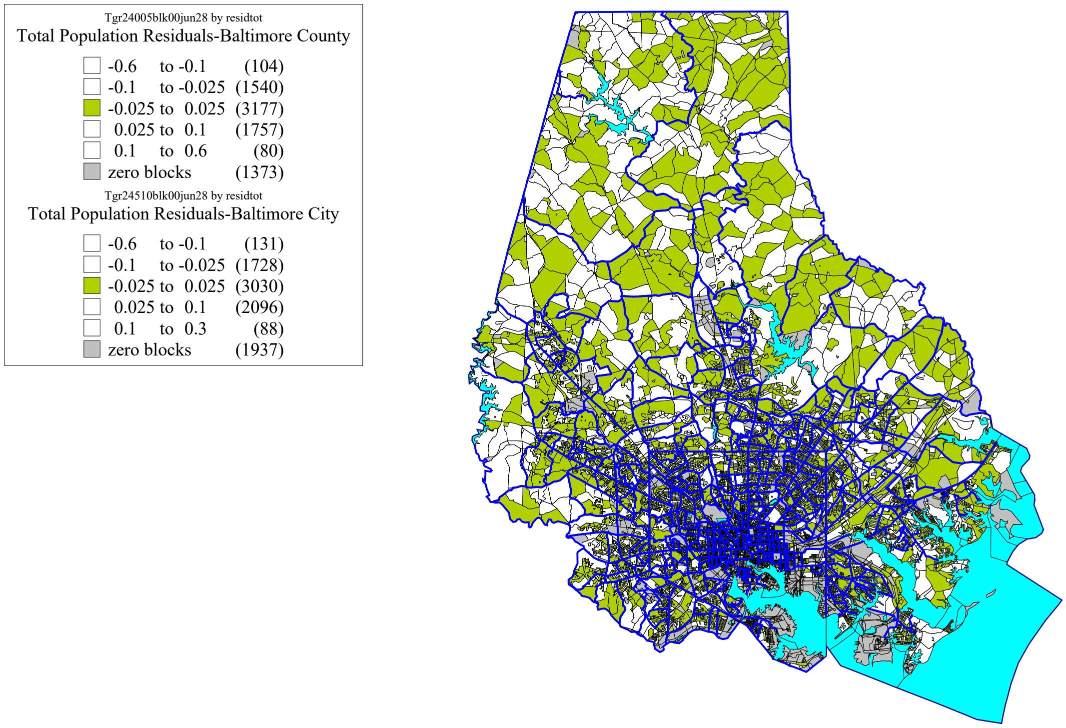


Figure 4.7.1b: Total Population ALPE Residuals-Downtown Baltimore City

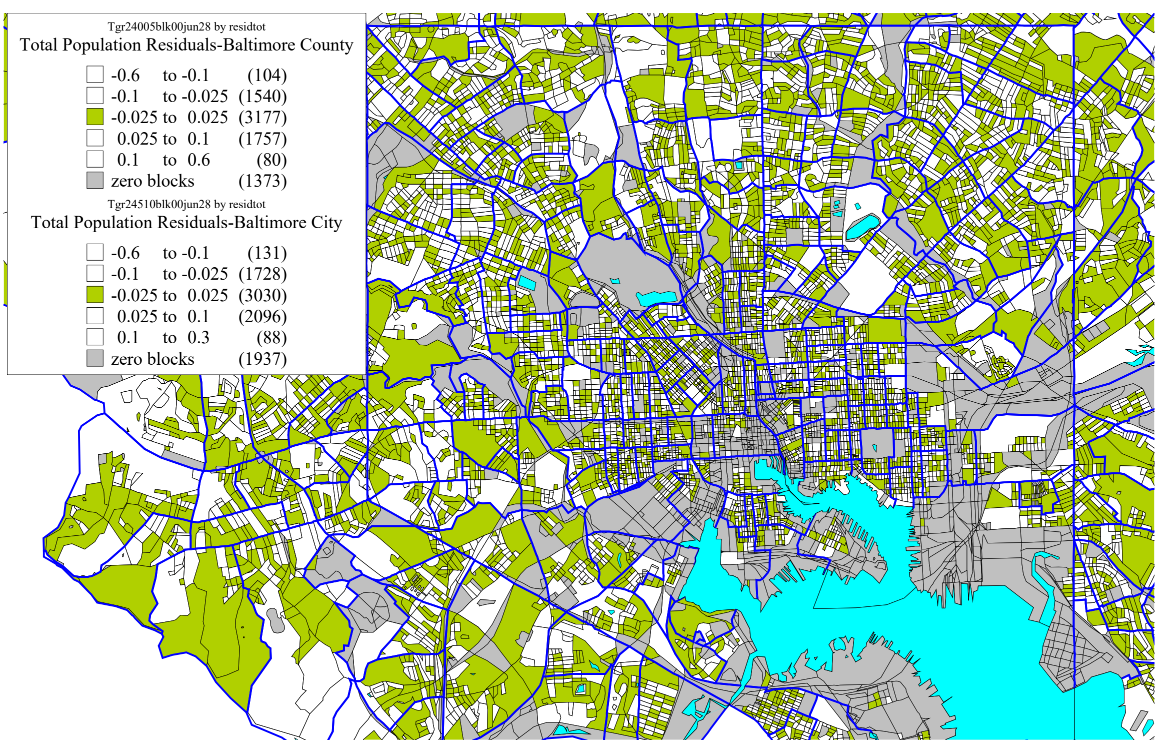


Figure 4.7.2a: Total Population ALPE Residuals-CO

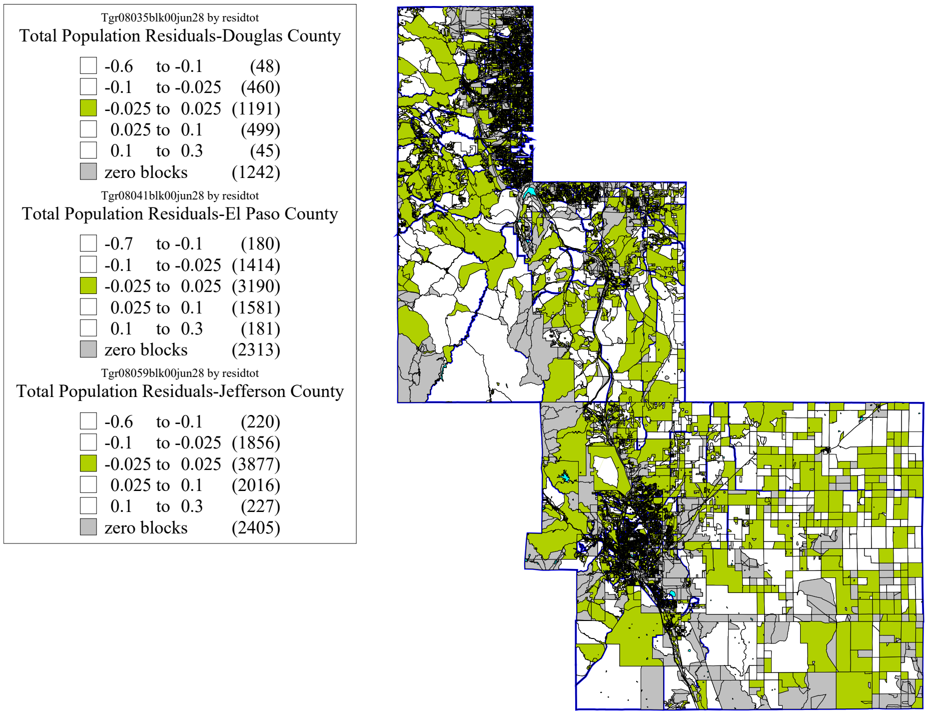


Figure 4.7.2b: Total Population ALPE Residuals-Downtown Denver

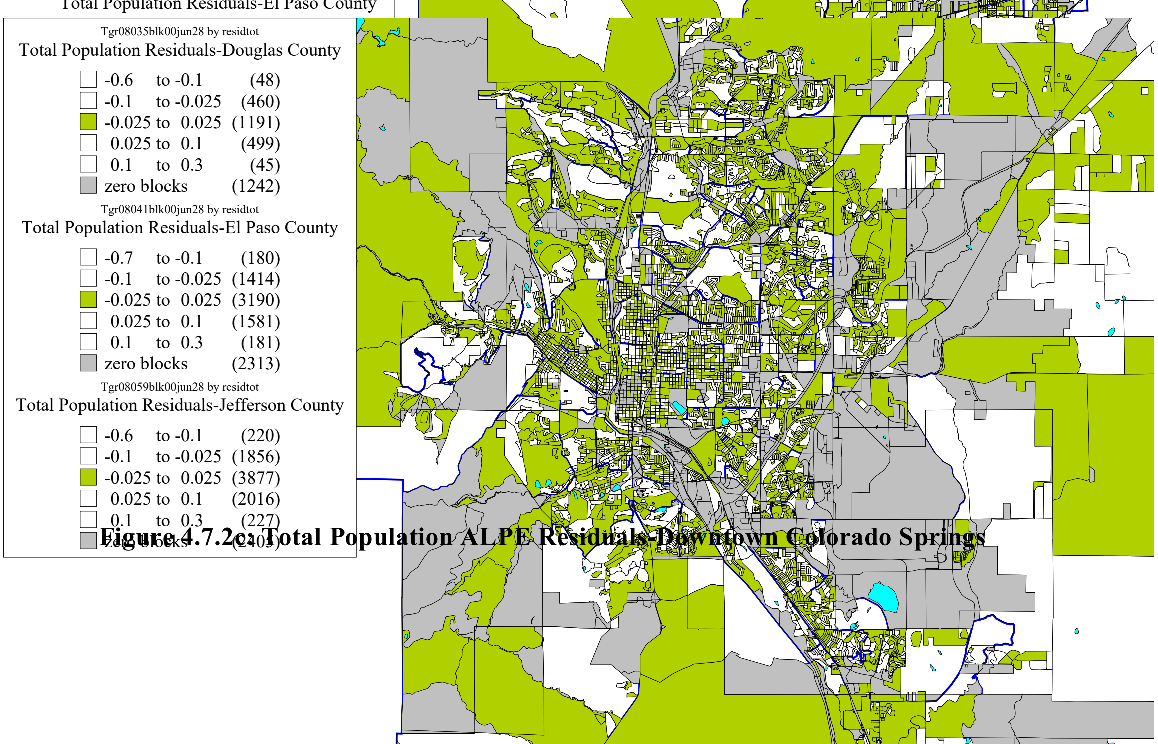


Figure 4.7.2c: Total Population ALPE Residuals-Downtown Colorado Springs

- Clustering of blocks with similar attributes (zero-blocks, large and small ALPEs) often occurs within tracts (thicker boundary) and appears to affect tract-level results.
- Most of the zero-blocks affecting total population ALPEs occur in commercial and industrial areas.

Figures 4.7.1a-b show the total population results for MD. Many of the small residuals tend to be clustered, indicating similarity between these adjacent blocks, and that statistical adjustment methods are more likely to be accurate in these areas. These clusters are apparent in Baltimore County and to a lesser extent in Baltimore City. The downtown blocks in Baltimore City exhibit a more random pattern of small vs. moderate/large residuals. There are two findings that may elaborate differences between block-level and tract-level results. First, it is not surprising to see that zero blocks (likely commercial and industrial areas) tend to be clustered within tracts. The zero blocks are concentrated in Baltimore City but are also present in several regions of Baltimore County. In some tracts, a large number of zero blocks are present (mixed commercial/ industrial/residential areas), and the tract-level ALPE results are less stable due to smaller denominators in ALPE calculations. This increases the differences between block- and tract-level distributions due to inflated tract-level results. That is, the denominator in the mean calculations goes from about 8,000-10,000 blocks to 200-400 tracts. And because of the clustering of small vs. moderate/large residuals, some tracts have large numbers of moderate/large residuals and suggest larger tract-level residuals.

The evidence is similar for the CO total population residuals (Figures 4.7.2a-c). Zero blocks are concentrated in urban areas and exist in several other regions. Small residuals also tend to be clustered. Visual inspection between the core urban areas of Denver and Colorado Springs does not indicate any differences between the cities. The CO findings also suggest that variability at the tract-level may be higher because of clustering of blocks and smaller denominators in tract-level calculations. However, no effort has been made to see if adjacent blocks with moderate/large ALPEs tend to offset each other, with similar numbers of positive and negative residuals.

Black ALPE residuals

Figure 4.7.3a: Black ALPE Residuals-MD

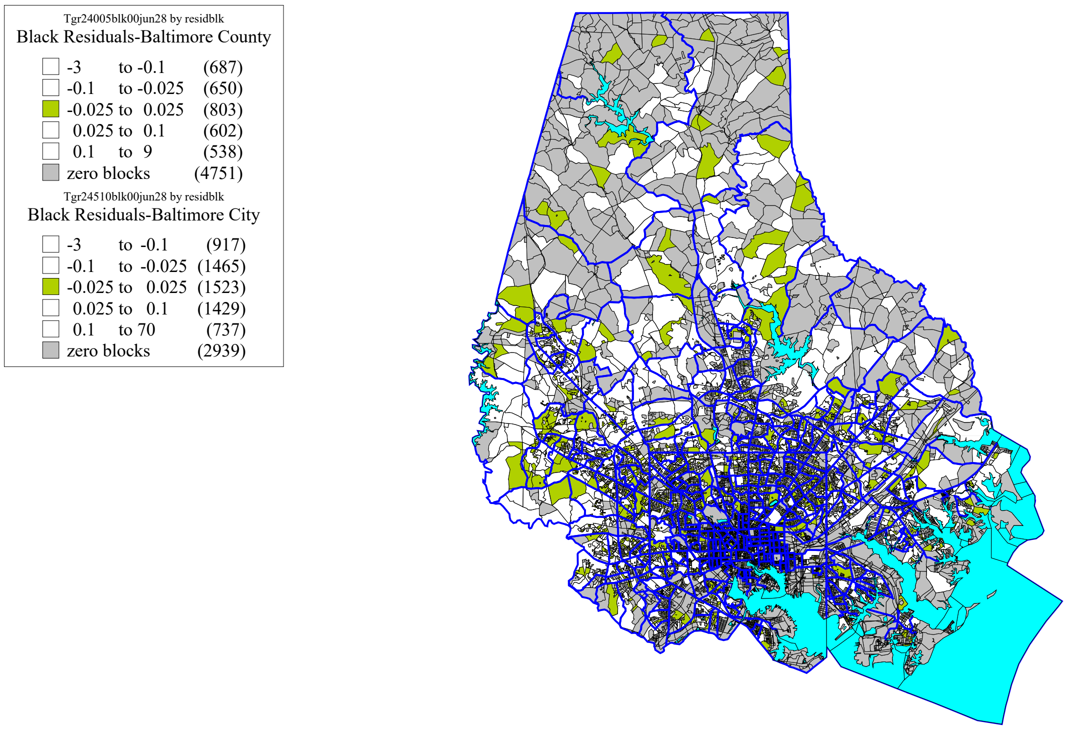
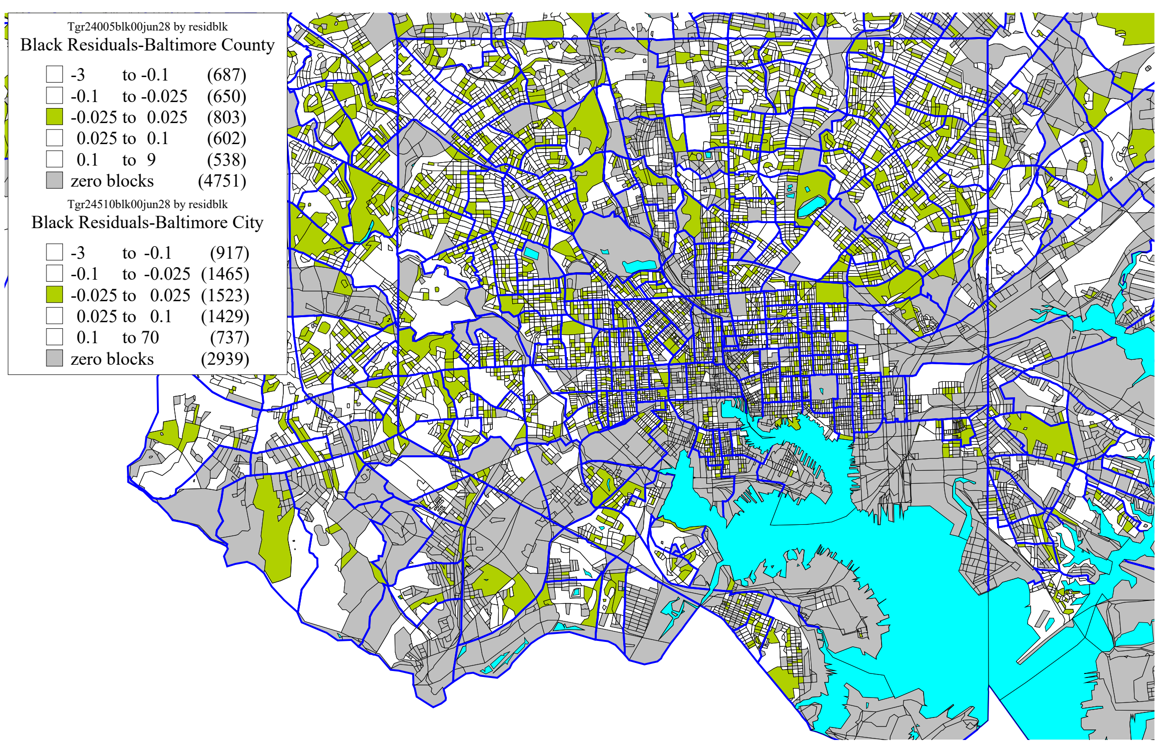


Figure 4.7.3b: Black ALPE Residuals-Downtown Baltimore City



- Black ALPE residuals are large and heterogeneous across Baltimore County due to segregated residential patterns.
- The distribution of tracts with large and small residuals is similar in Baltimore City and County, excluding the zero-blocks.
- Despite the greater number of blocks with Black residents in Baltimore City, a larger proportion of Baltimore City blocks has moderate residuals, compared to Baltimore County (48 percent vs. 38 percent).

Because the analysis of ALPE residuals for Blacks, Hispanics, and the age groups focuses on subsets of the total population, there are more zero blocks due to residential segregation patterns, compared to the total population maps. Census 2000 indicates that the majority of blocks in northern Baltimore County do not have Black residents. But the maps also suggest that a larger proportion of blocks have moderate and large ALPE residuals, compared to the total population ALPE residuals. This is supported in the distributional breakdown of ALPE residual categories in the map legend. These results also impact tract-level heterogeneity. In the northern Baltimore County tracts with few blocks having Black residents, ALPE residuals are likely to be large (if calculated). And due to the greater proportion of moderate/large ALPEs, potential tract-level ALPE residuals could also be more heterogeneous. The distribution of low vs. moderate/ large ALPE residuals is similar throughout Baltimore County, Baltimore City, and urban, downtown Baltimore.

Hispanic ALPE residuals

Figure 4.7.4a: Hispanic ALPE Residuals-CO

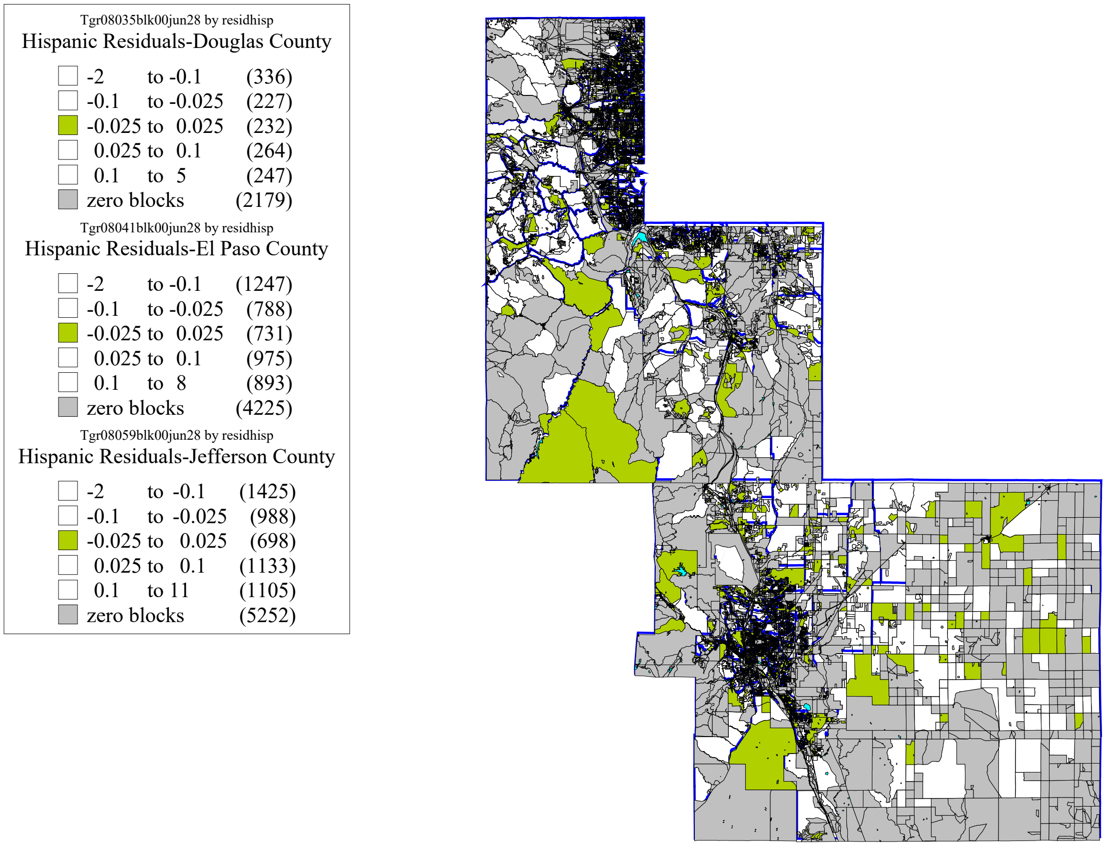


Figure 4.7.4b: Hispanic ALPE Residuals-Downtown Denver

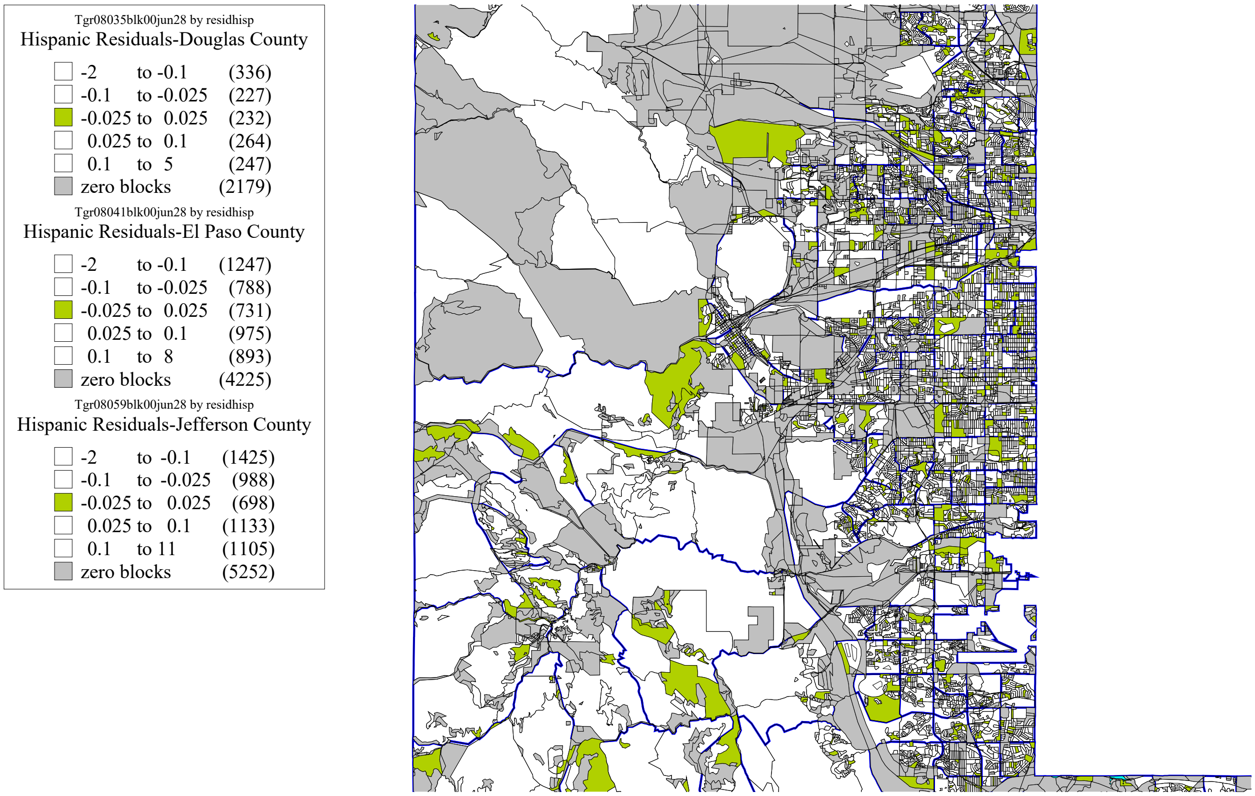
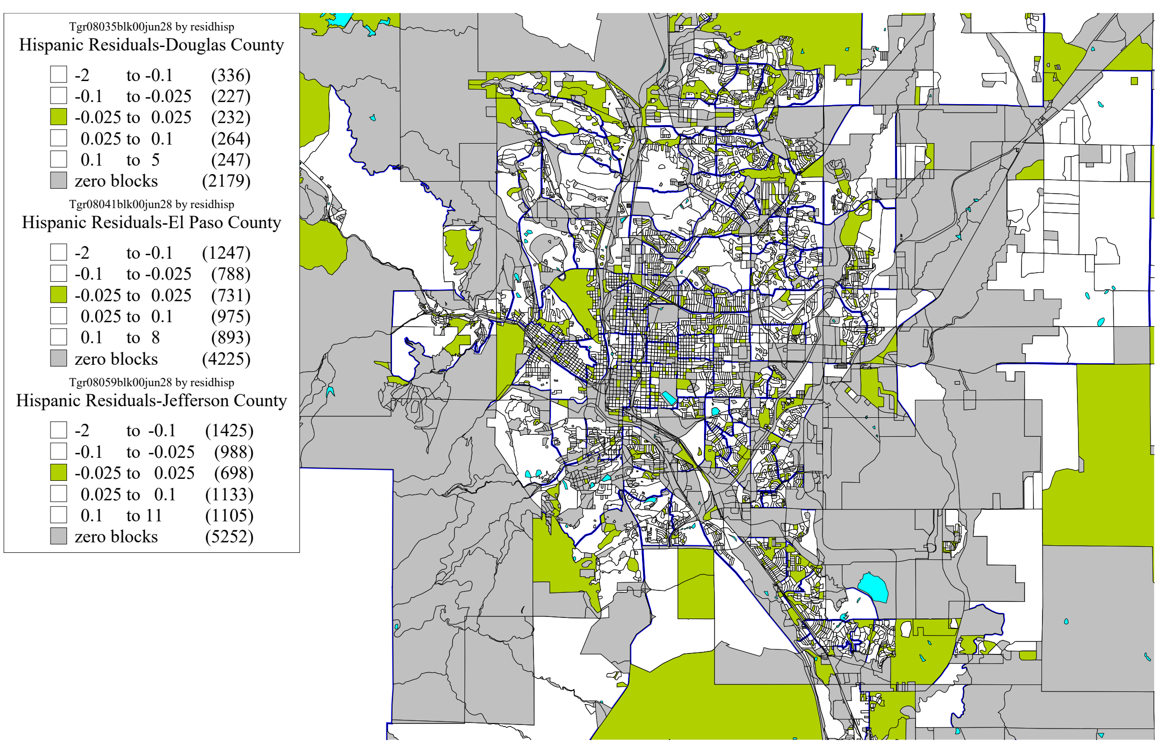


Figure 4.7.4c: Hispanic ALPE Residuals-Downtown Colorado Springs



- Hispanics tend to reside in urban rather than rural blocks.
- There is a similar proportion of blocks with large Hispanic residuals as there are large Black residuals in MD; however, there is a high proportion of large residuals in Douglas and El Paso counties.
- Hispanics tend to be clustered in urban areas and adjacent suburban areas of CO. A very small proportion of blocks has small ALPE residuals, and most tend to be moderate or large residuals across the three counties. The implication of Hispanic ALPE residuals is more extreme than found with Black ALPE residuals: tract-level errors are likely to have inflated denominators due to fewer blocks with Hispanic residents while tract-level errors are likely to be greater due to the smaller proportion of small Hispanic ALPE residuals. This is true in both urban and suburban areas.

5. RECOMMENDATIONS

This section synthesizes findings and describes immediate opportunities for improvement, areas of near-term further development, and directions for future research. Most of the recommendations are a direct result of this outcomes evaluation, but companion evaluation reports and staff discussions and opinions also influenced this list.

Immediate needs and opportunities for improving the accuracy and utility of administrative records:

5.1 Develop clear objectives, benchmarks for success, and timetables for accomplishing tasks. This issue is important for focusing the work of a limited staff and providing assurances that objectives are being met. Some decisions made during the computer processing or specifications phases can have unintended negative results. New methodologies and processes require careful evaluation over all phases of work. Ideally, multiple methods will be compared in test runs and the best overall choice selected for implementation. Are total population counts more important than demographic characteristics and should accurate tracts or block measures be the focus? What tolerance or level of error is acceptable for administrative records results? Should the immediate goal be accurate identification of individuals to improve linking with national surveys or are accurate tract-level characteristics more important? One strategy may include a consortium of federal agencies that would work with Census in an ongoing structured format to conceptualize needs, goals, and methods. A clear set of objectives would then facilitate the other recommended tasks for improving the accuracy and utility of administrative records.

5.2 Use the Bottom-Up enumeration method and separate household and group quarters populations. The Bottom-Up enumeration method produced more accurate household population counts for all counties. The address-matching process was important because it validated addresses found in administrative records. This led to unmatched addresses being replaced by actual Census results. These two activities were the most successful components in the administrative records processing. However, there needs to be further research on non-city-style addresses and how to identify corresponding physical addresses. Some addresses are the commercial mailing addresses of accountants, lawyers, guardians, and executors and not the physical addresses of actual persons in the administrative records. Address-related research should be expanded to improve the accuracy of block and tract population counts, as well as persons within households.

This evaluation demonstrated that administrative records provide accurate household population counts but ignored the group quarters population. Part of the strength of the Bottom-Up enumeration method is a reliance on accurate household addresses. Similarly, a transient population that resides in group quarters is unlikely to have consistent address records across administrative files, while lag time in processing administrative records for transient persons affects the accuracy of group quarters population counts. A separate process, through alternative administrative records, sampling or local surveys appears to be the choice for enumerating the group quarters population.

5.3 Obtain administrative records extracts that coincide with a specific day. A Census enumeration counts a population at an exact time and place. AREX processing was based on files that were collectively current for December, 1998 or Spring, 1999, but were compared to Census 2000. The direct consequence of this potential 15-month interval is that persons who died were reported in the administrative records, but not Census, while new births were reported in Census but not administrative records. Though many of the deficiencies described in this report are due to this 15-month interval, birth, death, and migration population counts may still be unreliable if files are poorly synchronized. And the address selection process also hinges on file consistency. For example, five files may have the same address but one does not. But address selection may ignore the sixth address because it is inconsistent, though perhaps more accurate than the address in the other five sources. Age distributions are affected by state policies on birth and death records and may not cover a specific place and time because of reporting lag issues. Finally, race and Hispanic origin distributions may be affected because migrants tend to be minorities and have higher fertility rates. Both sources of error (eliminating the lag interval and file synchronicity) can be quickly rectified through agency relationships and better planning.

5.4 Revise the race imputation methodology and discard model-based approaches to race imputation. The current race imputation model is perhaps the most deficient operation in the administrative records processing. Race information is seldom available for children because most federal agencies do not record these data. It is methodologically more difficult to impute race codes for individuals or small areas (including tracts and blocks), compared to counties and states. The reason for this is that model-based values reflect sample or aggregate characteristics and cannot provide the variability that occurs for individuals or small areas. Enhancing administrative records with Census 2000 may produce better results than previously available. However, about eight percent of respondents self-identified as multi-race or some other race and did not fit neatly into the five race categories. Annual births and deaths reflect about two percent of the U.S. population. And the effects of migration are not fully captured using the current race imputation methodology. Taken together, Census 2000 does not address all persons, while intercensal population changes need to be correctly enumerated.

Current race/ethnicity imputation methods rely on sample-based algorithms that apply mean values (based on subgroups) to individuals in AREX. Because national samples are used, the resultant mean values that are applied to individuals are frequently incorrect and result in inaccurate tract and block estimates of race and ethnicity. Methods that incorporate small area demographics that distinguish local vs. national mean values are necessary to improve small area estimates. But there is also significant unobserved heterogeneity that may occur, for example the surname list may be more accurate in some areas of the U.S. than others, or the children's imputation methodology may be affected by state policies that pass on demographic information to federal agencies. Census data appended to individual records should improve the accuracy of race assignments but may be less useful after 5-9 years. While Bayesian methods have the potential to improve the race imputation model, these methods require further development until they can be applied to small area analyses.

Areas of near-term further development

5.5 Develop alternative data sources and better methods for accurately counting births and deaths. AREX counts for the oldest and youngest persons suggest that birth and death information is not recorded in a timely manner. Further research is needed to understand whether this is due to the agency providing the data or delays prior to their receipt (i.e., other agencies, their processing schedules, and state regulations and policies). Births and deaths are recorded in administrative records after they are processed by county and state agencies. It's not clear how long the lag period is between an event and when it is recorded by federal agencies. Obtaining annual birth and death records from the National Center on Health Statistics (NCHS) also is affected by reporting lag. Obtaining records directly from states or from NCHS as it is received from states would minimize these lag intervals.

The demographic events of birth and death are extreme analogs to mobility because preceding and succeeding records do not exist. Births and deaths are local events that are administered by counties and states before processing at federal agencies. And because states may vary in the efficiency that they process data and their policies, regional variation in the accuracy of demographic events may exist in national files. This issue may be an aspect of the unobserved heterogeneity in the accuracy of young and old AREX individuals, impacting block and tract results. But because of the suddenness of these events, the impact of annual vs. frequently updated files becomes more important in identifying the most reliable source files for these age groups.

5.6 Obtain alternative data for identifying the race and ethnicity of children. Race and ethnicity generally comes from Social Security files that fail to document this information from birth certificates that were issued over the last 14 years. Additional data sources must be sought, possibly school enrollment data, though these data have been difficult to obtain. Accurate demographic characteristics of parents may carry over to children and resolve many of these missing race identifiers. But there are some problems with using parent information for children. Divorced and separated couples with dependent children may have less accurate parent information and could be placed at one physical address rather than another.

5.7 Further evaluate the use of administrative records for redistricting. Administrative records may provide an early source of data for redistricting and reapportionment as close total counts were achieved in most legislative districts. Administrative records provided reasonably good total population counts for most legislative districts, despite large AREX-Census differences in Census block totals.

State legislative districts are smaller than U.S. and state senate districts and are created by aggregating Census tabulation blocks. Despite large AREX-Census differences in Census block totals, AREX provided fairly accurate population counts for most districts. However, the age, race/ethnicity, sex characteristics of districts were not investigated. The findings of this study suggest that block count totals and the age/race distributions can be vastly improved in future administrative records databases and legislative districts will become even more accurate. This may allow redistricting and reapportionment efforts to commence early, reducing time constraints, while providing a greater opportunity for public review and comment on proposed boundaries.

5.8 Develop a new Hispanic name list. The race imputation process relies partly on surname lists to estimate the likelihood that an individual is Hispanic, Asian-Pacific Islander, or American Indian. While the Asian-Pacific Islander list was recently expanded using surnames from Census 2000, the Hispanic surname list requires similar updating. The surname lists are the only person-level identifiers of race and ethnicity outside of those recorded in the administrative records sources.

5.9 Research the address selection methodology. Current address selection methods have relied upon latest address date or most frequently recorded address. But posting dates may be the same across administrative files and more accurate in one or more files and less accurate than others. Further, there may be regional differences in the accuracy of addresses. For example, Alaska, Hawaii, Massachusetts, and Louisiana may have distinct address processing procedures and deadlines because they are not in the continental U.S. or lack typical county structures. Address selection processing should incorporate the validity of the different administrative files due to regional variations in the way they are processed.

Directions for future research

5.10 Study and document the internal specifications, methods, etc., of federal agency collectors of administrative data. This recommendation has been briefly mentioned in several immediate and near-term recommendations. There is a clear need to understand and document in detail the manner in which the various federal agencies collect their data to understand validity and reliability differences across files. This would allow 'grading' of data that could be used for weighting and comparing files. A second possibility is working with federal data collectors to change their collection methods in order to promote consistency across files.

5.11 Conduct additional research on transient subpopulations. Some of these issues were handled in Census 2000 through enumeration of special places and a group quarters census. Vacancy rates, type of tenure, presence of non-relative household members, and age/race/ethnic composition identify blocks that are more difficult to enumerate and require additional effort and resources. These factors may also be linked to non-response followup households that require special enumeration and imputation methods and include nursing home and hospital residents, and college-aged persons. College-age individuals are mobile due to their part-time residence at school and movement from dorms to temporary housing. But following school, they are also likely to relocate and later purchase a home, marry, and have children. It becomes problematic to identify the best address for persons in this age group, women may change their name, and children are born. Special attention needs to be focused on this age group because address and household changes are so tightly linked with each other.

5.12 Develop new methods for distinguishing blacks and whites when there is little information available. This problem may be resolved with a highly accurate method of identifying and/or imputing race. Ideally, using administrative records along with household and block/tract characteristics can be used to provide improved race measures. But there may still be problems, or race may be better identified in some regions than others. Alternative methods need to be researched that provide independent support for persons being white or black.

5.13 Identify strengths and weaknesses in using the MAF for administrative records. The Census MAF is being used as a 'gold standard' for identifying whether administrative records addresses are correct or not. But the administrative records may capture new construction starts sooner than the MAF. Or there could be unknown errors and deficiencies. It is important for subsequent research and processing to identify the strengths and weaknesses in the MAF to fortify the enumeration process using administrative records.

REFERENCES

Aguirre International (1995). *Public Concerns About the Use of Administrative Records*. Unpublished document available from the US Census Bureau, July 12, 1995.

Alvey, Wendy and Scheuren, Fritz (1982). Background for an Administrative Record Census. *Proceedings of the Social Statistics Section*, Washington, DC: American Statistical Association, 1982.

American Statistical Association (1977). "Report of the Ad Hoc Committee on Privacy and Confidentiality," *The American Statistician*, *31*: 59-78.

Bell, Patricia (1999). Racial/Ethnic Homogeneity of Neighborhoods and Variation in Census Coverage of African Americans (unpublished manuscript)

Brackstone, G.J. (1988). "Statistical Uses of Administrative Data: Issues and Challenges," in Coombs and Singh, eds., *Statistical Uses of Administrative Data : An International Symposium*, Ottawa: Statistics Canada.

Buser, Pascal, Huang, Elizabeth, Kim, Jay K., and Marquis, Kent (1998). *1996 Community Census Administrative Records File Evaluation*. Administrative Records Memorandum Series # 17. Washington, DC: U.S. Bureau of the Census.

Bye, Barry (1997). *Administrative Record Census for 2010 Design Proposal*. Washington, DC: US Department of Commerce.

Coombs, J.W. and Singh, M.P. (1988). *Statistical Uses of Administrative Data : An International Symposium*. Ottawa: Statistics Canada.

Czajka, John L., Moreno, Lorenzo, Schirm, Allen L. (1997). *On the Feasibility of Using Internal Revenue Service Records to Count the U.S. Population*. Washington, DC: Internal Revenue Service.

Duncan, Otis Dudley and Duncan, Beverly (1955). A methodological analysis of segregation indices. *American Sociological Review*, *20*: 210-217.

Edmonston, Barry and Schultze, Charles (1995) *Modernizing the U.S. Census*, Washington DC: National Academy Press.

Flippen, Chenoa and Tienda, Marta (2000). Pathways to Retirement: Patterns of labor Force Participation and Labor Market Exit Among the Pre-Retirement Population by Race, Hispanic Origin, and Sex. *Journal of Gerontology: Social Sciences*, 55B:1, S14-27.

Gellman, Robert (1997). *Report on the Census Bureau Privacy Panel Discussion*. Unpublished document available from the US Census Bureau, June 20, 1997.

Hayward, Mark D. and Melonie Heron. 1999. Racial Inequality in Active Life Among Adult Americans. *Demography*, 36:1, 77-91.

Huang, Elizabeth and Kim, Jay (2000). *One Percent Sample Study Report*. Administrative Records Research Memorandum Series #42, US Census Bureau.

Jenks, George F. and Caspall, Fred G. (1971). Error on Choropleth Maps: Definition, Measurement, Reduction. *Annals of the Association of American Geographers*, 61: 217-44.

Knott, Joseph J. (1991). *Administrative Records*. Memorandum for Distribution List, Bureau of the Census, Washington DC, November 12, 1991.

Krebs, C.J. (1989). *Ecological Methodology*. New York, NY: Harper and Row.

Myrskyla, Pekka (1991). Census by questionnaire--Census by registers and administrative records: The experience of Finland. *Journal of Official Statistics*, *7*: 457-474.

Myrskyla, Pekka, Taeuber, Cynthia, and Knott, Joseph (1996). *Uses of administrative records for statistical purposes: Finland and the United States*. Unpublished document available from the US Census Bureau.

Neugebauer, Sharon (1996). *Administrative Records File Acquisition History for the 1995 Census Test*. 1995 Census Test Results, Memorandum No. 24. US Census Bureau.

Pistiner, Arona and Shaw, Kevin A. (2000). *Program Master Plan for the Census 2000 Administrative Records Experiment (AREX 2000)*. Administrative Records Research Memorandum Series #49. US Census Bureau.

Plane, David A. and Rogerson, Peter A. (1994). *The Geographical Analysis of Population: With Applications to Planning and Business*. New York, NY: John Wiley and Sons.

Prevost, Ron (1997). *The Usefulness of IRS Information Returns in the Development of a National Administrative Records Database*. Administrative Records Research Memorandum Series #12, US Census Bureau.

Rogerson, Peter A. (2001). *Statistical Methods for Geography*. Thousand Oaks, CA: Sage Publications, Inc.

Sailer, Peter, Weber, Michael, Yau, E. (1993). *How Well Can IRS Count the Population?* Proceedings, Government Statistics Section, American Statistical Association. Alexandria, VA: American Statistical Association.

Shryock, H.S. and Siegel, J.S. (1973). *The Methods and Materials of Demography*. Washington DC: US Bureau of the Census.

Singer, Eleanor and Miller, Esther (1992). R*eactions to the use of Administrative Records: Results of Focus Group Discussions*. Census Bureau report, Center for Survey Methods Research, August 24, 1992.

Smith, Stanley K. and Shahidullah, Mohammed (1993). Evaluating Population Projection Errors for Census Tracts (paper presented at the annual meeting of the Population Association of America, April 1-2, 1993).

Sweet, Elizabeth (1997). *Using Administrative Record Persons in the 1996 Community Census*. Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Taueber, Cynthia, Lane, Julia, and Stevens, David (2000). *The Why, What, and How of Converting Program Records and Summarized Survey Data to State and Community Information Systems*. Paper presented at the Conference, Developing Public Policy Applications with Summarized Survey Data and Community Administrative Records. Baltimore, MD, June 6-7, 2000.

Weidman, Lynn and Alexander, Charles (1999). *Estimation for the American Community Survey: Ongoing work, planned work, and issues*. Paper presented to the Census Advisory Committee of Professional Associations Meeting, October 21-22, 1999.

Wurdeman, Kent (1996). *1995 Coverage Study-Phase II Results*. 1995 Census Test Results, Memorandum No. 52. US Census Bureau.

Zanutto, E. (1996). *Estimating a population roster from an incomplete census using mailback questionnaires, administrative records, and sampled nonresponse followup*. Presentation to the US Bureau of the Census, August 26, 1996.

Zanutto, Elaine and Zaslavsky, Alan M. (1996). *Estimating a population roster from an incomplete census using mailback questionnaires, administrative records, and sampled nonresponse followup*. In Proceedings of the US Bureau of the Census Annual Research Conference. Washington, DC: US Census Bureau.

Zanutto, Elaine and Zaslavsky, Alan M. (1996). *Modeling census mailback questionnaires, administrative records, and sampled nonresponse followup, to impute census nonrespondents*. In Proceedings, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Zanutto, Elaine and Zaslavsky, Alan M. (2001). *Using administrative records to impute for nonresponse*. To appear in R. Groves, R.J.A. Little, and J.Eltinge (eds.), Survey Nonresponse. New York, NY: John Wiley.

Administrative Records Experiment in 2000 (AREX 2000) Outcomes Evaluation

APPENDICES

APPENDIX 1: PROFILE OF TEST SITES

A1.1 County demographics

The ability to accurately measure the resident population using administrative records is likely to vary by the age, race, sex, and Hispanic composition of the AREX counties. These demographic groups are likely to have distinct coverage rates within administrative records, as well as mobility, fertility, and mortality rates. The latter rates are also likely to interact with the record-keeping processes of the federal agencies that collect and maintain the data. The sites were chosen for their varying demographic characteristics to test the feasibility of enumerating the population using administrative records. Table A1.1 provides a detailed breakdown of 2000 demographic characteristics for the five counties in the AREX test sites. Some general comments on the AREX test sites include:

- Baltimore and Baltimore City have the largest populations, compared to the less populated CO counties.
- Females exceed males in all five counties; the sex ratio is larger in the CO counties.
- The MD counties are much older than the CO counties; the age 0-4 age group proportions are larger in the CO counties, while the older age groups are larger in the MD counties.
- Baltimore City, and to a lesser extent, Baltimore County, have large Black populations; Hispanics are the largest minority population in CO, followed by APIs.

Table A1: Demographic Breakdown of the Census 2000 Household Population for AREX Counties

	Baltimore County		Baltimore City		Douglas County		El Paso County		Jefferson County	
Total	736,652		625,401		175,300		501,533		519,326	
White	548,776	74.5%	196,427	31.4%	162,639	92.8%	408,167	81.4%	471,107	90.7%
Black	147,226	20.0%	404,198	64.6%	1,663	0.9%	31,875	6.4%	4,126	0.8%
AI	1,923	0.3%	2,097	0.3%	716	0.4%	4,725	0.9%	3,971	0.8%
API	23,631	3.2%	9,168	1.5%	4,488	2.6%	13,954	2.8%	12,330	2.4%
Hispanic	13,433	1.8%	10,712	1.7%	8,825	5.0%	56,677	11.3%	51,346	9.9%
Age 0-4	45,179	6.1%	41,593	6.7%	16,949	9.7%	39,006	7.8%	33,213	6.4%
5-19	147,393	20.0%	135,558	21.7%	41,376	23.6%	115,404	23.0%	111,655	21.5%
20-24	41,740	5.7%	43,627	7.0%	5,478	3.1%	32,596	6.5%	28,901	5.6%
25-34	100,363	13.6%	89,525	14.3%	28,552	16.3%	75,205	15.0%	70,672	13.6%
35-44	122,116	16.6%	97,983	15.7%	38,007	21.7%	90,039	18.0%	96,357	18.6%
45-54	107,499	14.6%	81,691	13.1%	26,235	15.0%	68,878	13.7%	84,174	16.2%
55-64	67,187	9.1%	53,630	8.6%	11,597	6.6%	37,709	7.5%	46,190	8.9%
65+	105,175	14.3%	81,794	13.1%	7,106	4.1%	42,696	8.5%	48,164	9.3%
65-74	54,768	7.4%	43,533	7.0%	4,784	2.7%	24,988	5.0%	28,025	5.4%
75-84	40,114	5.4%	29,618	4.7%	1,959	1.1%	14,211	2.8%	15,900	3.1%
85+	10,293	1.4%	8,643	1.4%	363	0.2%	3,497	0.7%	4,239	0.8%
Male	349,319	47.4%	288,070	46.1%	87,478	49.9%	248,764	49.6%	257,876	49.7%
Female	387,333	52.6%	337,331	53.9%	87,822	50.1%	252,769	50.4%	261,450	50.3%

A1.2 Spatial and ecological issues affecting AREX tracts

Summary: Though it appears that tracts with moderate/high population density have more vacant and/or rental units, this is not true for all tracts in the MD and CO AREX counties. Some higher density tracts may have more desirable neighborhoods and fewer vacant units. Similarly, there is evidence that suburban and rural tracts may have less stable net migration of residents. In some cases, new home construction may be related to vacant units, however, the spatial maps do not identify new home subdivisions.

Figure A1.1a: Number of Vacant Housing Units: MD Tracts

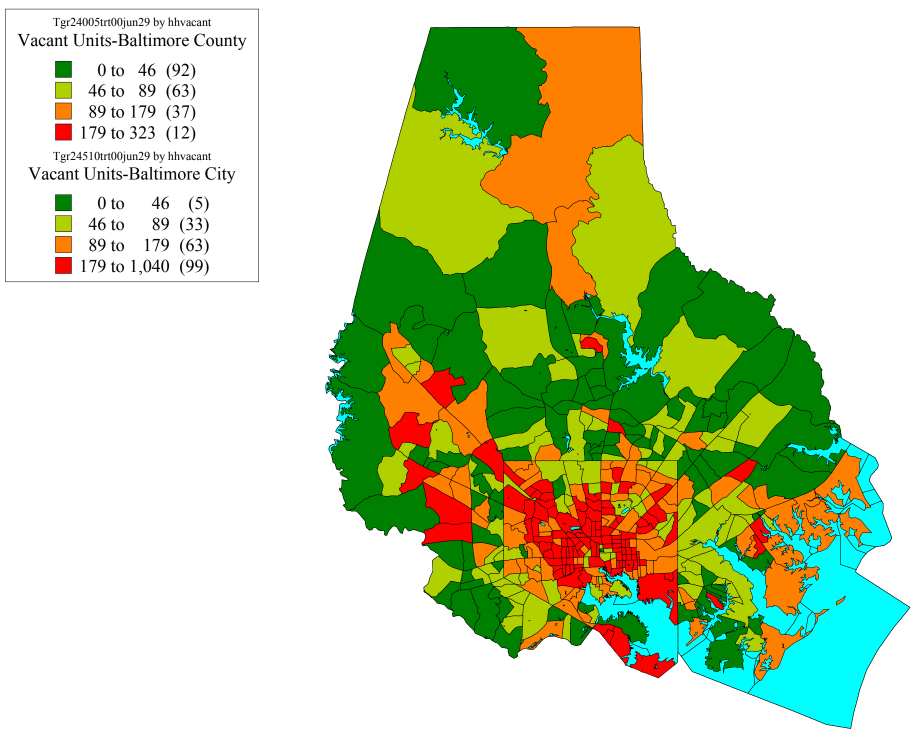


Figure A1.1b: Number of Vacant Housing Units: CO Tracts

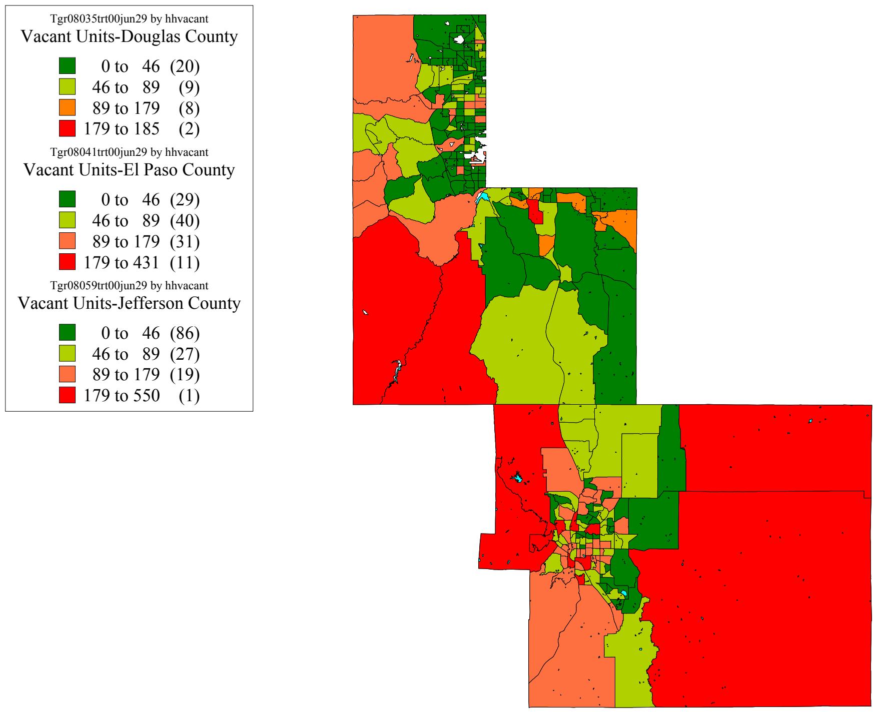


Figure A1.2a: Population Density: MD Tracts

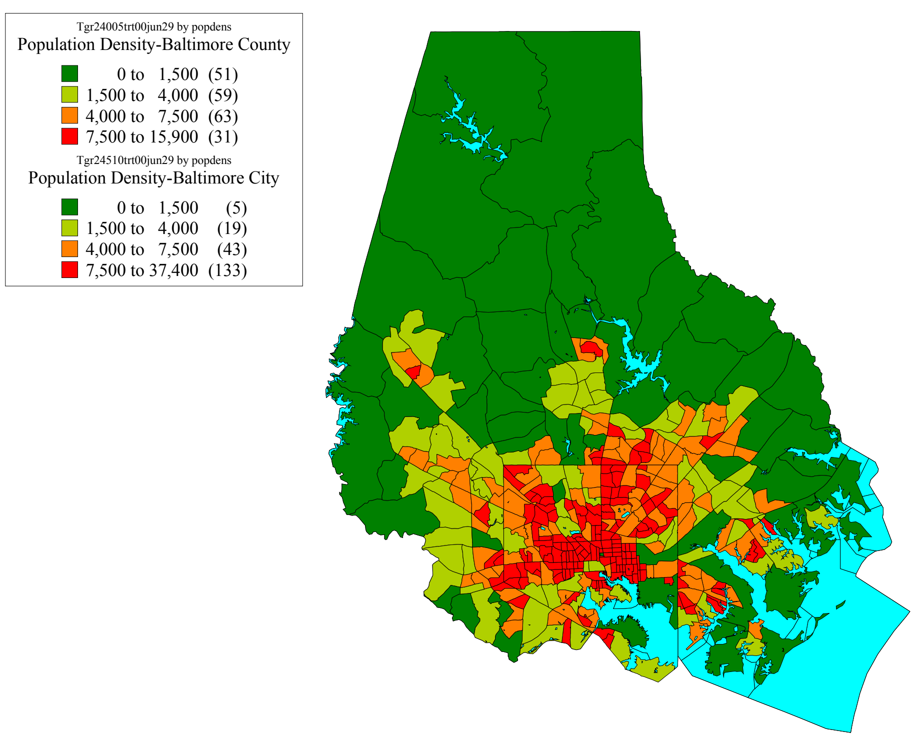
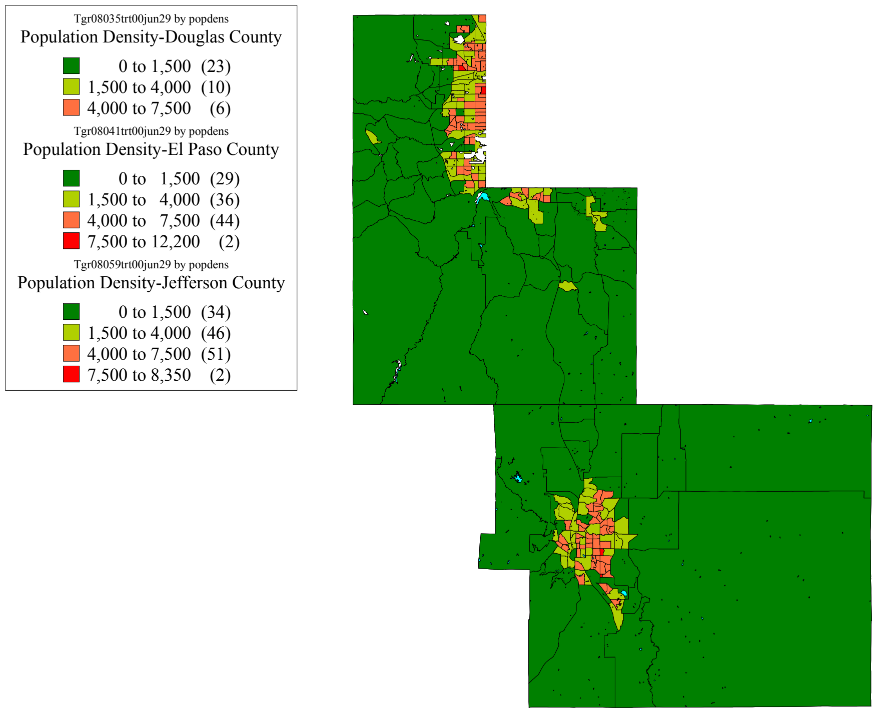


Figure A1.2b: Population Density: CO Tracts



Figures A1.1a-b, A1.2a-b show the ecological distribution of vacant housing units and population density of MD and CO tracts from Census 2000 results. These basic ecological maps suggest that tracts with larger proportions of vacant and/or rental (high-density) units are heterogeneously distributed across the AREX counties. Some of the highlights include:

- Tracts with large numbers of vacant units coincide with high-density population tracts, though this is not true for all tracts, especially around Denver in Jefferson County.
- A large number of tracts have vacant housing units, especially in downtown Baltimore City, with several tracts having clusters of moderate and high numbers of vacancies in Baltimore County.
- Despite the large land area of the CO tracts, there are few tracts with large numbers of vacant housing units; most of the vacant units are in El Paso county, within and around Colorado Springs.

Spatial and ecological issues impact how well administrative records accurately measure the resident populations of sub-county regions and their proximity to each other, and can have a variable affect on demographic group counts. Counties with a large number of vacant housing units are likely to provide poorer estimates because of the reporting lag between a moving household and federal agencies recording of population mobility. Residents of these areas may be less affluent and potentially less-covered populations. Similarly, transient population groups, like college students and military personnel, can flow into and out of other residences and group quarters. Older residents, and especially women, are more likely to enter or exit nursing homes, compared to the general population. This group also experiences higher mortality rates that may impact their coverage, due to reporting lag in recording mobility or deaths.

A1.3 Demographic diversity of AREX tracts

Summary: Age diversity is greater in urban and suburban tracts of MD, while race/Hispanic diversity is greater in urban and suburban tracts of CO. The Black population in Baltimore City is highly segregated and appears to be as homogeneous as mostly White tracts in the other counties. Some tract counts are harder to measure accurately, particularly those where multi-race reporting occurs and large numbers of non-relative household members live (not shown). These harder to measure attributes tend to affect the same tracts.

Figure A1.3a: Shannon-Wiener Diversity Index for Age-MD Tracts

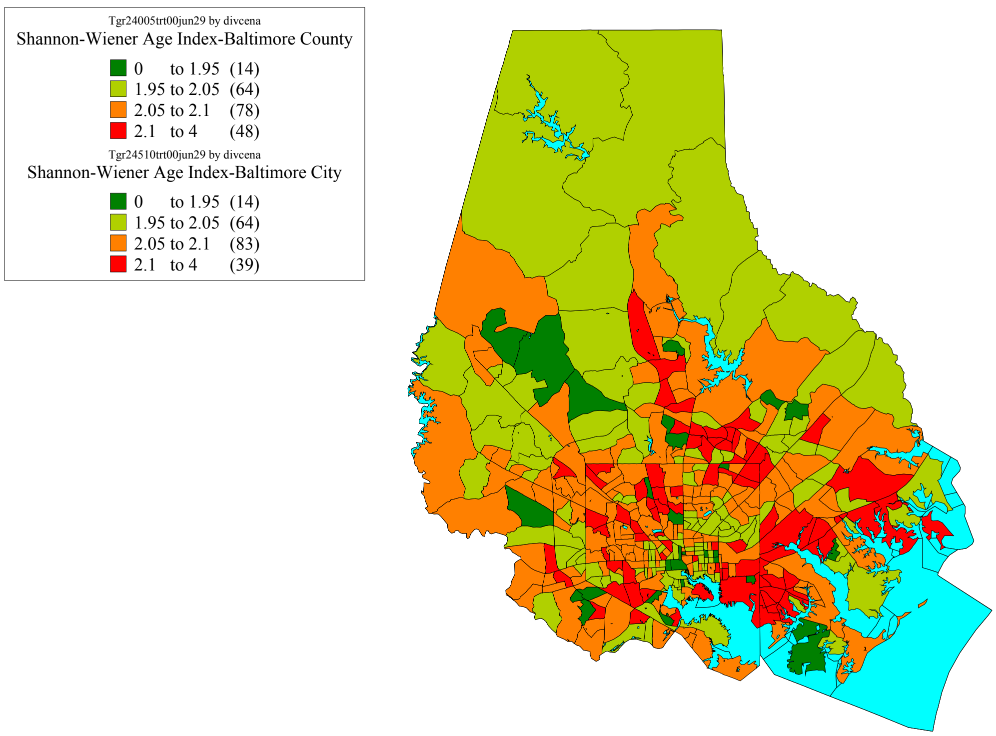


Figure A1.3b: Shannon-Wiener Diversity Index for Age-CO Tracts

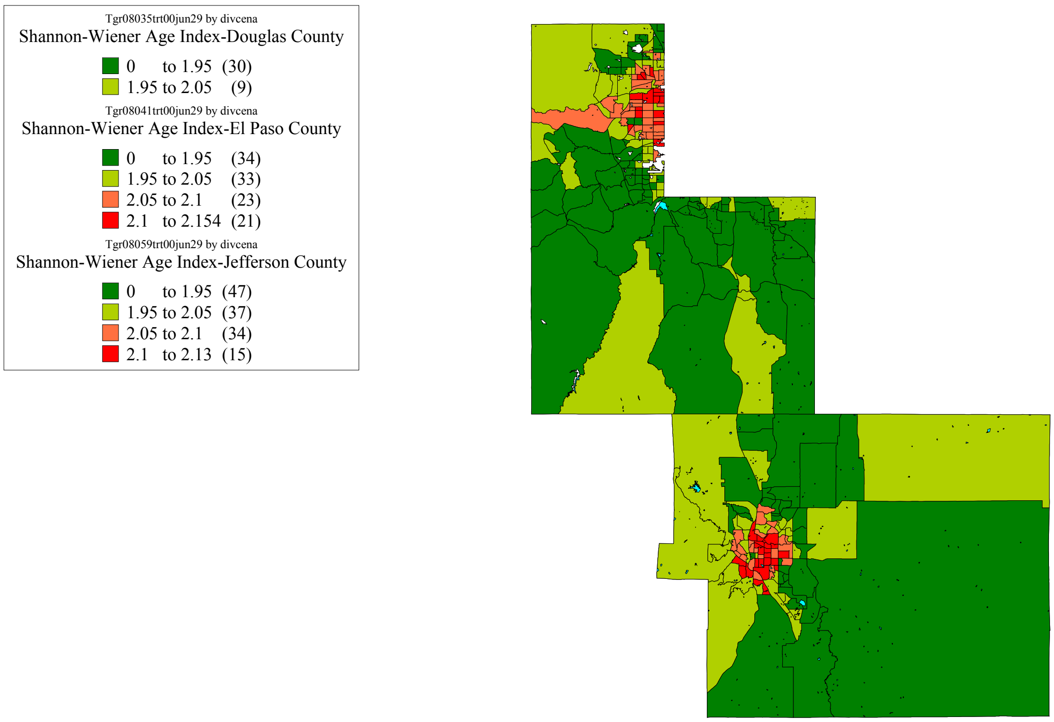


Figure A1.4a: Shannon-Wiener Diversity Index for Race-MD Tracts

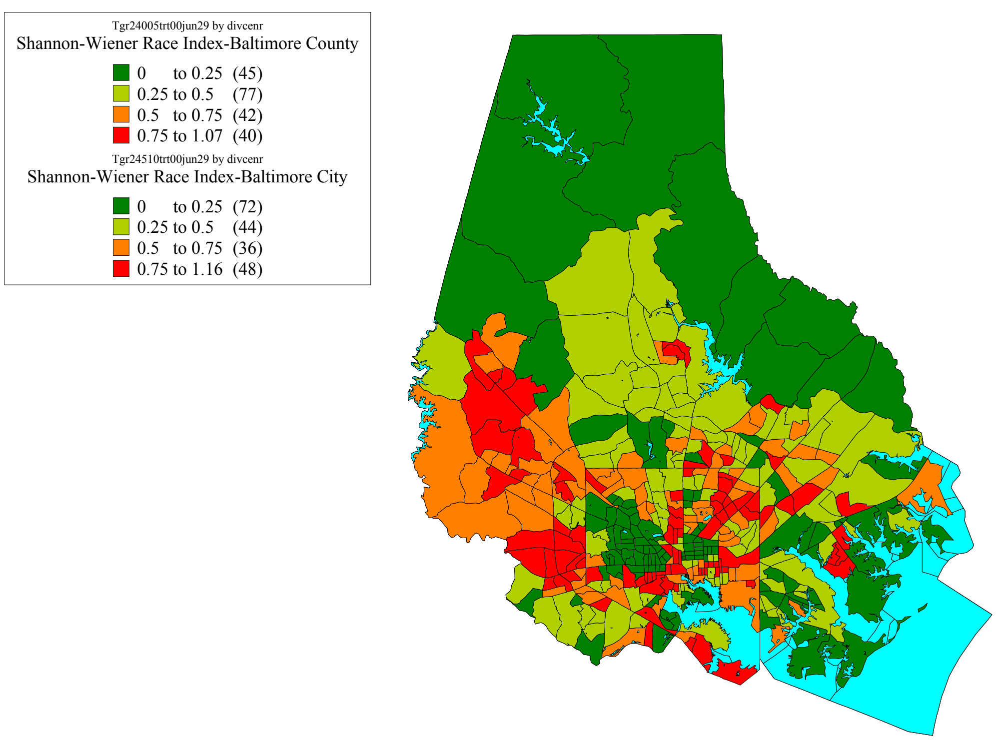
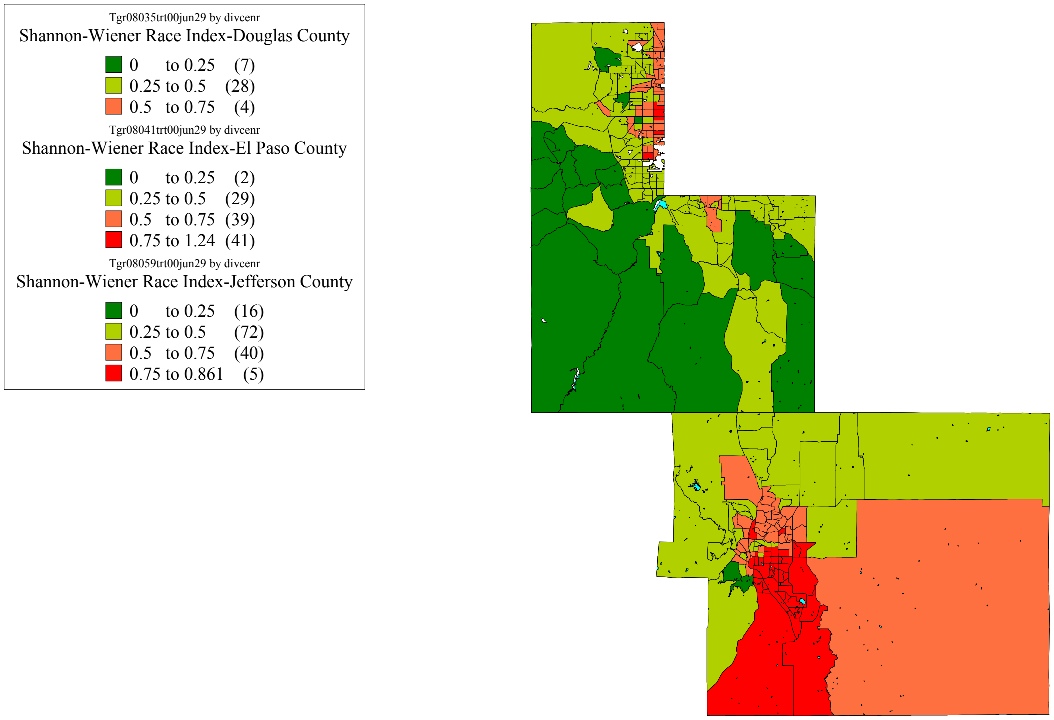


Figure A1.4b: Shannon-Wiener Diversity Index for Race-CO Tracts



The demographic characteristics of tract residents and the type, price, and availability of housing units are likely to attract or repel new in-migrants and affect tract-level coverage rates. The Shannon-Wiener diversity index measures the number of race/Hispanic groups and their population proportions within a tract, but does not distinguish whether a tract is predominantly White or Black. Tract-level diversity using Census 2000 results is shown above in Figures A1.3a-b, A1.4a-b and can be summarized:

- In the MD counties, the most diverse tracts exist in the southern, more urban section of MD; the western portion of Baltimore City with a large proportion of Blacks appears to be as racially uniform as the White, northern portion of Baltimore County.
- In the CO counties, diversity is concentrated in urban areas and several bordering tracts; this pattern may also reflect tracts with a large proportion of Hispanics and smaller White minority.

APPENDIX 2: RACE IMPUTATION

General description of the race imputation process

The race imputation process used logistic model results estimated from linked CPS-SSA Numident files, as well as Hispanic and Asian surname files and IHS records (see Bye, 1998 for complete details). The general model algorithm used the Numident, IHS, and surname identifiers to predict the matched CPS race codes. The type of Numident record, frequency of consistent race reports, geographic identifiers, and foreign birth indicators were also for calculating race probabilities. The calculated probabilities were then processed through a hot deck procedure for the final race assignment.

Persons under the age of 18 frequently lacked complete information and had blank race assignments in their Numident records. More problematic is that CPS did not include persons under age 15 years and the original model results did not address this younger age group. Consequently, the race information was incomplete and potentially inaccurate for minor children and a second stage imputation process was applied. The derived race assignment of the primary tax filer was applied to all children. While this second stage may address problems with children's records, it may also assign race from inaccurate race identifiers of some householders.

Table A2 provides the results of the race assignment process and imputed race codes by type of assignment:

Table A2: Race Assignment and Imputation Rates by Method, Race, and County

Imputation Method Most Frequent Report[1]	Baltimore County	Baltimore	Douglas	El Paso	Jefferson
All Persons	81.0%	74.8%	69.0%	72.9%	75.8%
White	82.5%	75.2%	40.2%	74.6%	77.6%
Black	79.9%	75.8%	61.3%	77.2%	59.5%
AI	55.7%	54.1%	35.5%	38.3%	36.3%
API	64.4%	56.4%	56.7%	64.5%	61.2%
Hispanic	1.2%	2.1%	1.8%	5.5%	4.4%
Imputed Primary Tax Filer Race (applied to persons under 18)[2]					
All Persons	9.4%	7.9%	13.1%	10.1%	9.7%
White	9.1%	6.7%	13.3%	10.2%	9.8%
Black	10.8%	8.6%	13.2%	12.0%	10.1%
AI	8.3%	7.5%	11.2%	9.9%	9.0%
API	8.8%	4.9%	10.3%	9.2%	9.7%
Hispanic	-	-	-	-	-

PCF Probability Model (applied to all adults)[2]

All Persons	3.1%	1.8%	4.1%	6.8%	6.3%
White	2.9%	4.3%	3.6%	6.7%	5.7%
Black	0.9%	0.3%	13.6%	3.3%	21.5%
AI	20.5%	16.7%	15.1%	10.3%	12.1%
API	19.9%	15.8%	17.9%	18.1%	21.5%
Hispanic	92.5%	82.6%	84.6%	85.3%	88.2%

[1]Most frequent race report / total AREX records
[2]Imputed records / total AREX records

APPENDIX 3: TRACT AND BLOCK INCONGRUITIES

Technical factors affecting tract and block differences

The relationship between level of geography and the accuracy of AREX counts is more complicated than it appears. For total population counts, county-level results can be hypothesized as more accurate than tract-level results, which in turn are expected to be more accurate than block-level results. And this relationship was supported by total population values across the geographic levels. However, statistical, computational, and substantive issues affect this relationship when looking at sparse populations that are likely to be distributed in a heterogeneous fashion across counties.

Table A3.1 (next page) is a listing of blocks for a single tract that focuses on AI residents and indicated AREX overcounted Census by 250 percent.[1] Each record shows the block level Algebraic Percent Error (ALPE) and AREX and census counts and difference for that block. This single tract covers 34 blocks, but only three have AI residents, based on Census results, while AREX indicates one block has AI residents. However, there are four blocks with AI residents, according to AREX, but three are zero-blocks for Census. Because of the computational problems, the block level results have two blocks each with 100 percent undercounts of census. But the five AREX persons who were not counted at the block-level contributed to a 267 percent overcount at the tract-level (11-3)/3.[2]

There is reason to be skeptical about the validity of the AREX overcounts for Census zero blocks. AREX overcounts may indicate a single person in a block is an AI but one would expect at least two or three AIs in a block, reflecting family members and neighbors with similar backgrounds living in the same neighborhood. The validity of these overcounts is important when considering the accuracy of the various geographic levels. One would expect the greatest accuracy at the county-level, because AREX overcounts could be 'absorbed' by the larger population counts. At the tract level, AREX overcounts are included in calculations, but tract-level denominators are sometimes small, resulting in inflated ALPE overcounts and highly skewed distributions that are sometimes U-shaped. At the block-level, AREX overcounts are not included in the distributions and calculations because the zero-blocks render these as undefined. This is problematic for small populations and sparse distributions, especially AIs and persons 75+ or 85+.

[1] Actual tract numbers have been dummied to ensure confidentiality.
[2] This ALPE exceeds the 95th percentile and was topcoded to 2.5.

Table A3. Block Counts of American Indians for a Sample Tract

Block	Tract Blks/tract	ALPE	Block ALPE	****AI Block counts***** AREX	Census	Difference
1234501.47	34	2.5	.	0	0	0
1234502.47	34	2.5	-1	0	1	-1
1234503.47	34	2.5	.	0	0	0
1234504.47	34	2.5	.	0	0	0
1234505.47	34	2.5	.	0	0	0
1234506.47	34	2.5	.	0	0	0
1234507.47	34	2.5	-1	0	1	-1
1234508.47	34	2.5	.	0	0	0
1234509.47	34	2.5	.	0	0	0
1234510.47	34	2.5	.	.	0	.
1234511.47	34	2.5	.	0	0	0
1234512.47	34	2.5	.	0	0	0
1234513.47	34	2.5	.	1	0	1
1234514.47	34	2.5	.	0	0	0
1234515.47	34	2.5	.	0	0	0
1234516.47	34	2.5	.	0	0	0
1234517.47	34	2.5	.	0	0	0
1234518.47	34	2.5	.	0	0	0
1234519.47	34	2.5	.	0	0	0
1234520.47	34	2.5	.	0	0	0
1234521.47	34	2.5	.	0	0	0
1234522.47	34	2.5	.	0	0	0
1234523.47	34	2.5	.	0	0	0
1234524.47	34	2.5	.	0	0	0
1234525.47	34	2.5	.	0	0	0
1234526.47	34	2.5	.	0	0	0
1234527.47	34	2.5	.	0	0	0
1234528.47	34	2.5	.	0	0	0
1234529.47	34	2.5	.	0	0	0
1234530.47	34	2.5	.	0	0	0
1234531.47	34	2.5	.	4	0	4
1234532.47	34	2.5	0.000	1	1	0
1234533.47	34	2.5	.	5	0	5
1234534.47	34	2.5	.	0	0	0
Tract Total				**11**	**3**	**8**

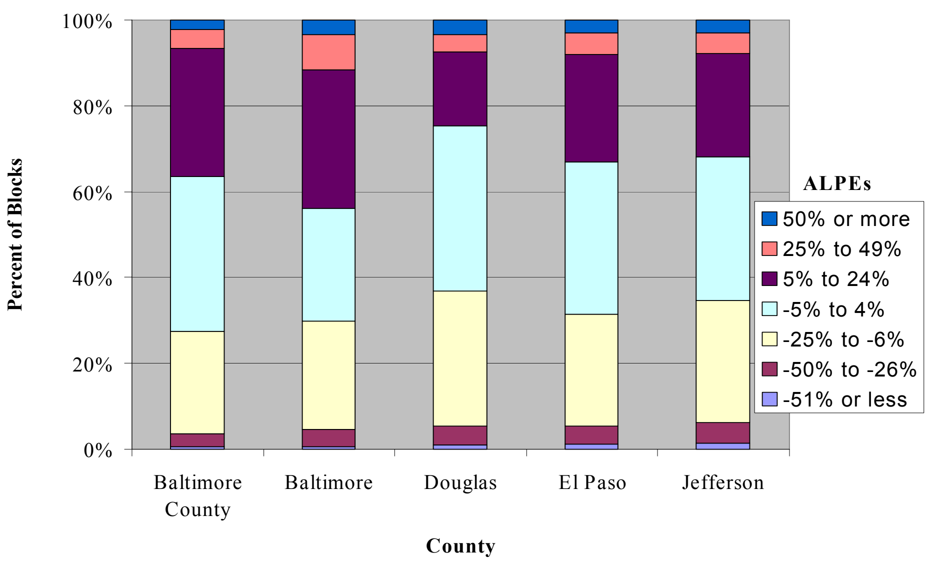
APPENDIX 4: BLOCK-LEVEL ANALYSES

Block-level demographic ALPEs

Summary of Results: The block-level ALPE results provided the least accurate measure of total population (26 to 38 percent of blocks met the five percent criterion and about 85 percent met the 25 percent criterion), compared to tract and county results. But block results were better than tract ALPEs for sex and selected age groups (0-4, 20-24, 65+, older age groups). Race groups with larger populations provided better estimates of Census counts at the five percent criterion, but all block-level ALPEs were worse using the 25 percent criterion. The block-level results exclude zero blocks and mean county ALPEs are affected by smaller denominators, an especially important issue for small population groups that reside in few blocks.

(Figure repeated from section 4.4)

Figure 4.4.1: Distribution of Blocks with Under- and Overcounts of Total Population



The block ALPE results describe the accuracy of counts at the smallest geographic level and relative to counties and tracts. The main problem with this type of comparison is the ALPE denominator potentially inflates block-level ALPEs for small population subgroups and especially minorities. This inflation is likely to be greater than found in the tract-county comparisons. A second issue affecting comparisons is the exclusion of blocks where census did not identify persons with a particular attribute (zero blocks). Tract and block ALPEs include blocks with zero counts because these blocks were collapsed into larger geographies. However, the block-level ALPEs use the reduced sample of blocks and the results may be quite different when comparing the ALPEs at various geographies.

TOTAL POPULATION

- AREX was more accurate in estimating tracts than blocks in all counties; from 26 to 38 percent of blocks were within the five percent criterion, and about 85 percent were within the 25 percent criterion in the five counties; Douglas County had the best results at the five percent criterion and Baltimore County was best at the 25 percent criterion.
- In the MD counties, slightly more blocks had moderate or large overcounts (ALPEs exceeding five percent, compared to the CO counties where more blocks had moderate undercounts (minus five percent to –24 percent; distributions not shown).

The AREX counts were less accurate at the block-level. Total population proportions are likely to be less accurate at smaller areas due to incorrect assignment of households at tracts and blocks that average out for county-level counts. This is demonstrated by the greater number of moderate and large ALPEs and indicates how smaller denominators and AREX processing flaws influenced the results. Though zero blocks were excluded and fewer blocks met the five percent criterion, a surprisingly large proportion of blocks met the 25 percent criterion in all five counties.

SEX

Figure A4.1a: Proportion of Blocks With Sex ALPEs Below 5% and 25% - Baltimore County

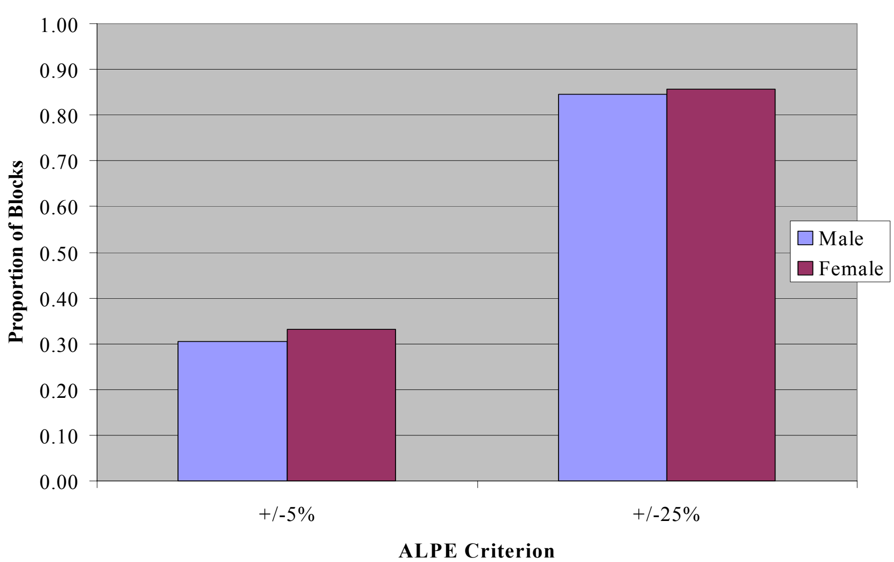


Figure A4.1b: Proportion of Blocks With Sex ALPEs Below 5% and 25% - Baltimore City

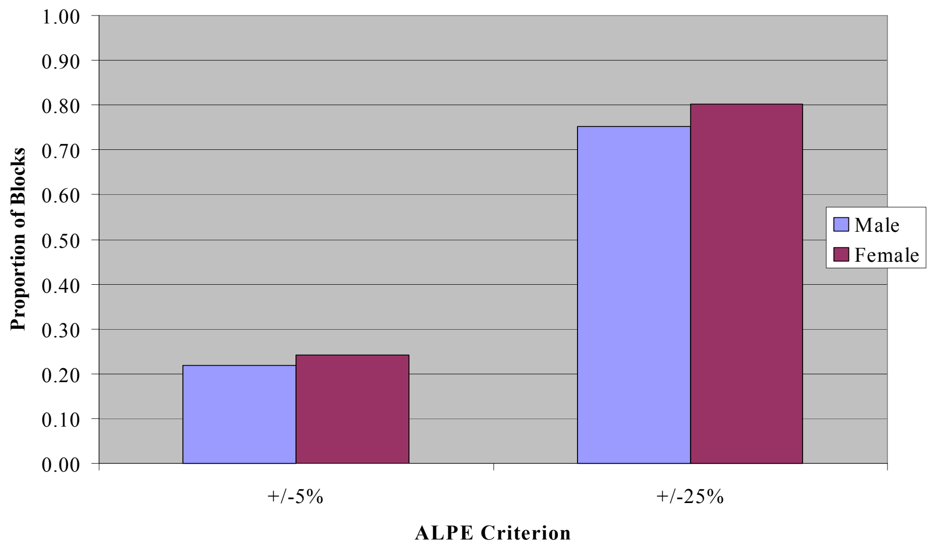


Figure A4.1c: Proportion of Blocks With Sex ALPEs Below 5% and 25% - Douglas County

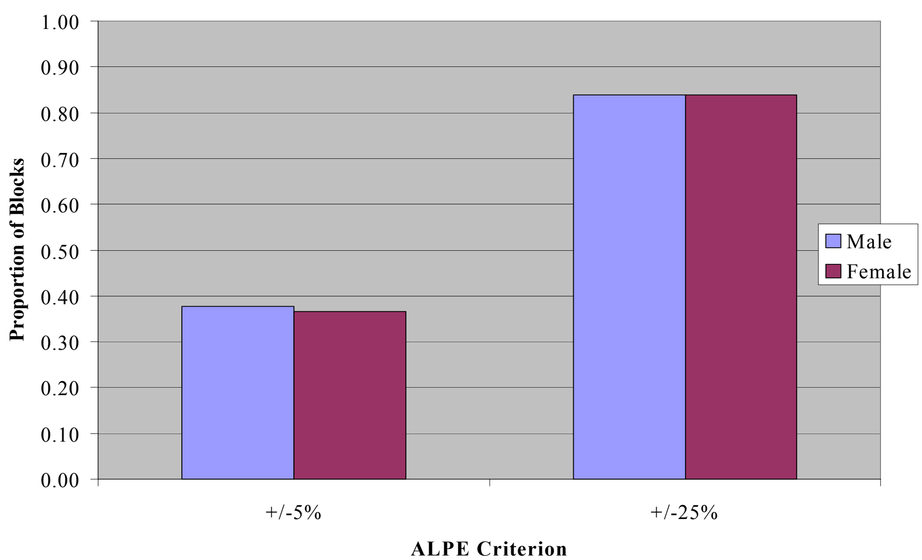


Figure A4.1d: Proportion of Blocks With Sex ALPEs Below 5% and 25% - El Paso County

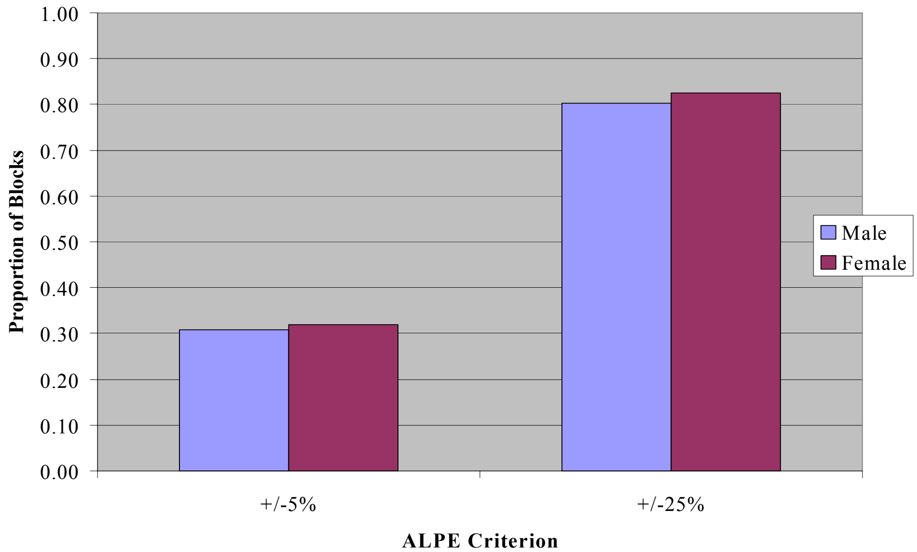
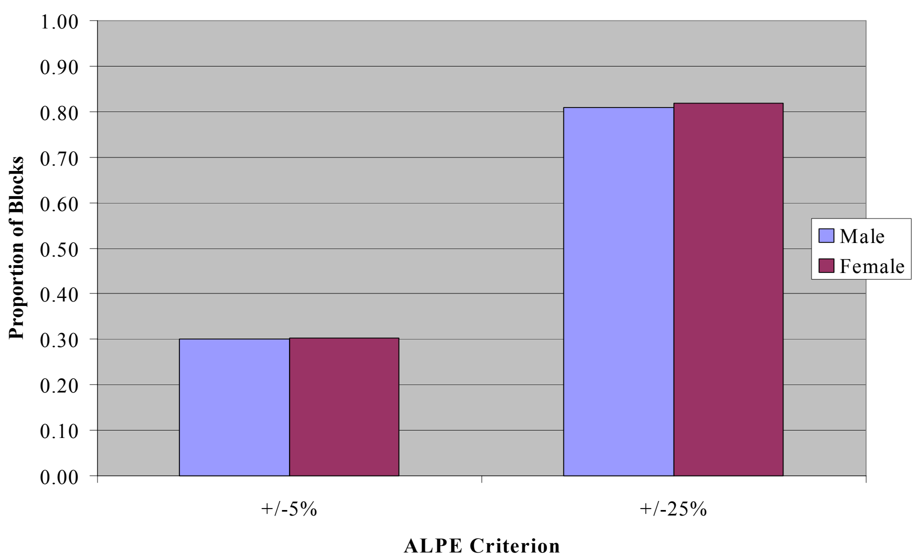


Figure A4.1e: Proportion of Blocks With Sex ALPEs Below 5% and 25% - Jefferson County



- The accuracy of AREX sex results at the five percent criterion was better for blocks than tracts.
- From 39 to 55 percent of male and female ALPEs were within the five percent criterion in the five counties; from 91 to 94 percent of blocks were within the 25 percent criterion.

Male and female undercounts were similar at all geographic levels and reflected the total population results. This similarity suggests that AREX processing was neutral towards whether individuals were male or female. However, males and females have different demographic rates (migration and mortality) at different points in the life-cycle, which may account for the small differences in the male and female AREX results.

AGE

Figure A4.2a: Proportion of Blocks with Age ALPEs Below 5% and 25% - Baltimore County

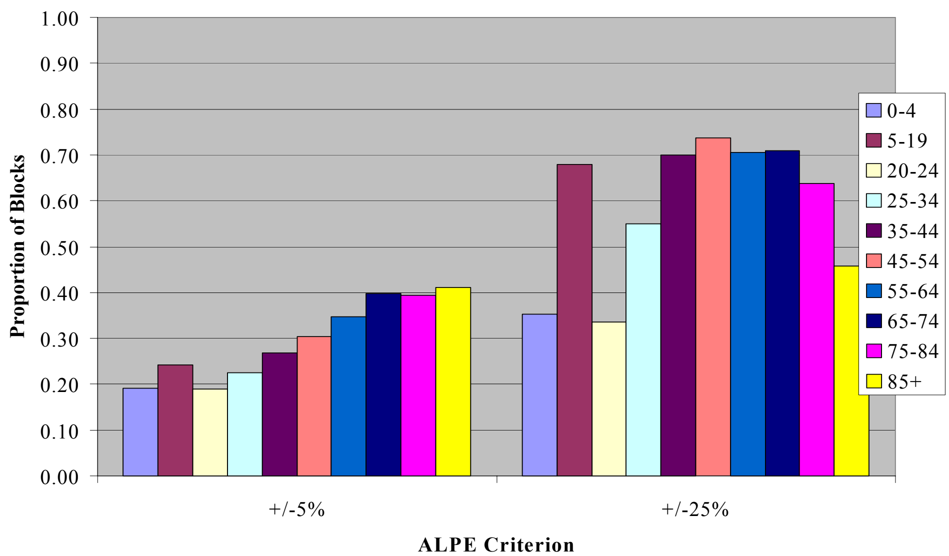


Figure A4.2b: Proportion of Blocks with Age ALPEs Below 5% and 25% - Baltimore City

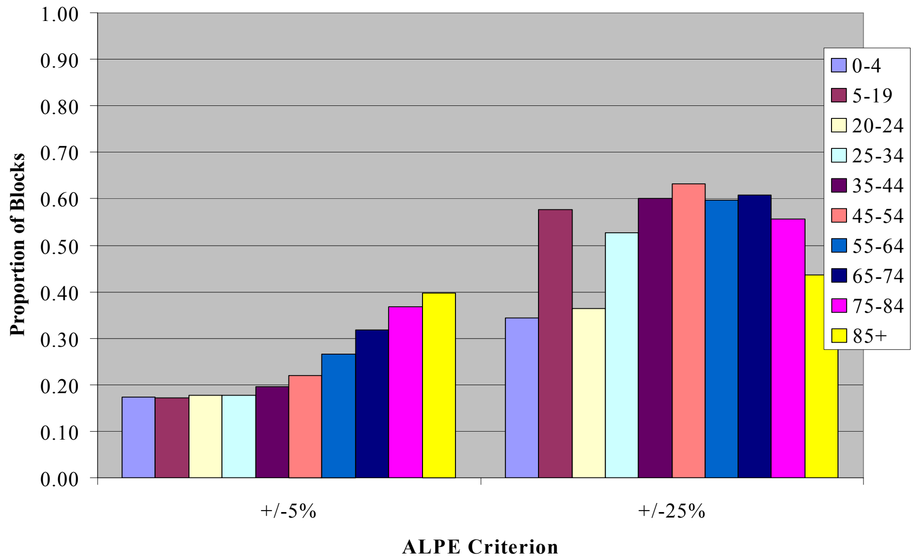


Figure A4.2c: Proportion of Blocks with Age ALPEs Below 5% and 25% - Douglas County

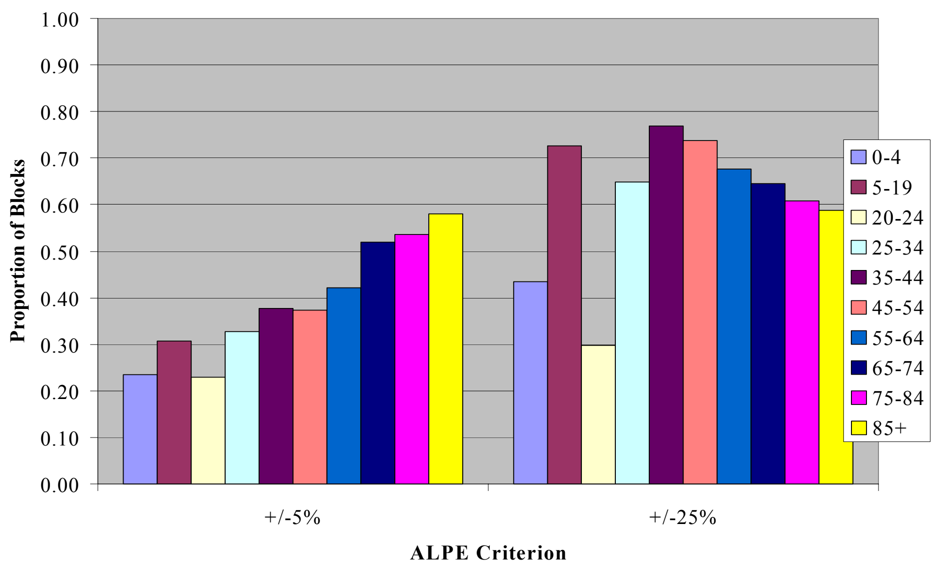


Figure A4.2d: Proportion of Blocks with Age ALPEs Below 5% and 25% - El Paso County

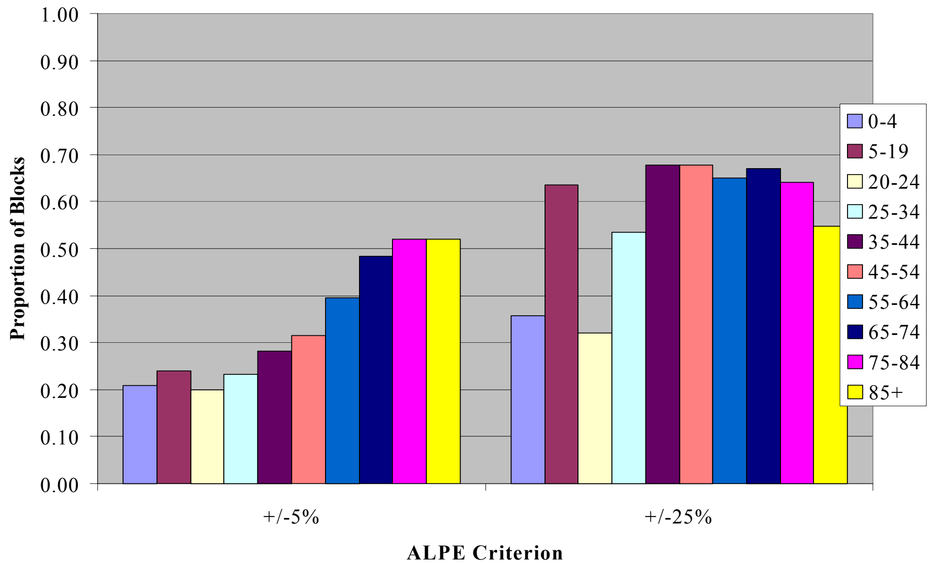
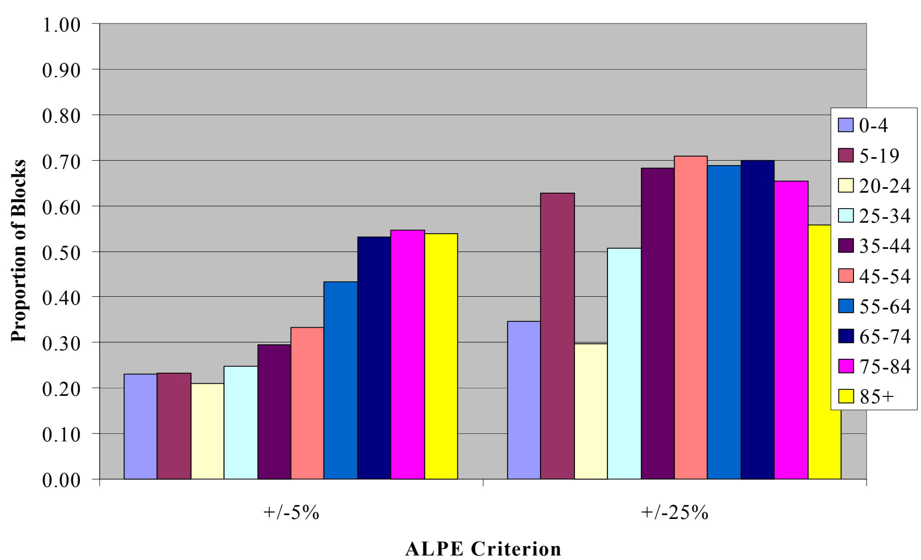


Figure A4.2e: Proportion of Blocks with Age ALPEs Below 5% and 25% - Jefferson County



- Age ALPE results support previous findings from tract and county results: AREX counts were within five percent of Census counts more often for the age 25-74 groups than younger age groups.
- The age ALPE results for age 25-64 age groups were much worse for blocks than tracts in all counties at both five percent and 25 percent criteria; however, block-level results were better for the age 0-4, 20-24, and 65+ age groups at the five percent criterion.
- Old age ALPEs at the five percent criterion were much better for blocks than tracts; though a smaller proportion of blocks had ALPEs of less than five percent, compared to tracts; results for the 75-84 and 85+ age groups were as good or better than for the 65-74 age group.

In general, the block-level results for age were less accurate than the tract-level ALPE results. Besides having smaller denominators for ALPE calculations, blocks with zero population counts are excluded from the analyses. But if AREX performs poorly in some blocks and those blocks are contiguous, it suggests that some block-level ALPE results may be better than corresponding tract ALPEs. That is, errors may be smaller in blocks but cumulated into larger ALPEs within tracts. This may be the case for the 0-4, 20-24, and 65+ age groups because a larger proportion of blocks (compared to tracts) met the five percent criterion.

RACE / ETHNICITY

Figure A4.3a: Proportion of Blocks with Race ALPEs Below 5% and 25% - Baltimore County

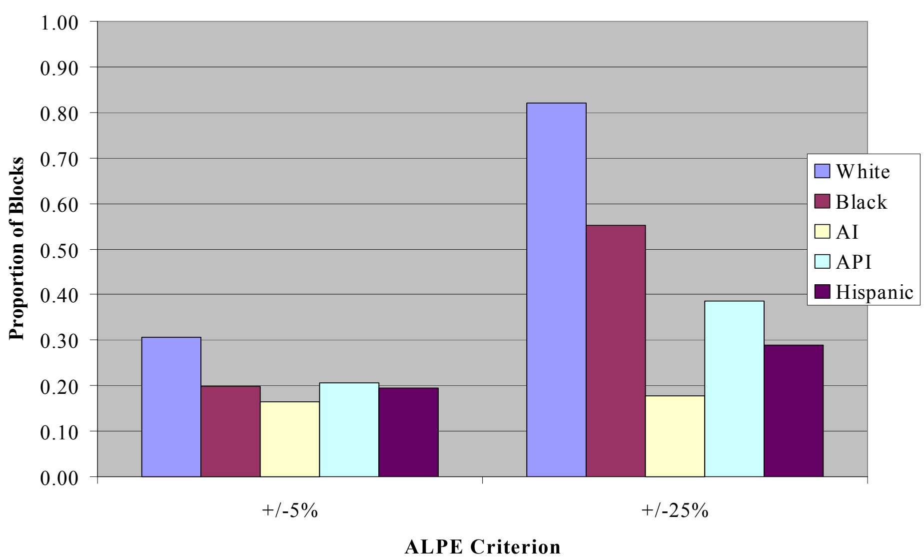


Figure A4.3b: Proportion of Blocks with Race ALPEs Below 5% and 25% - Baltimore City

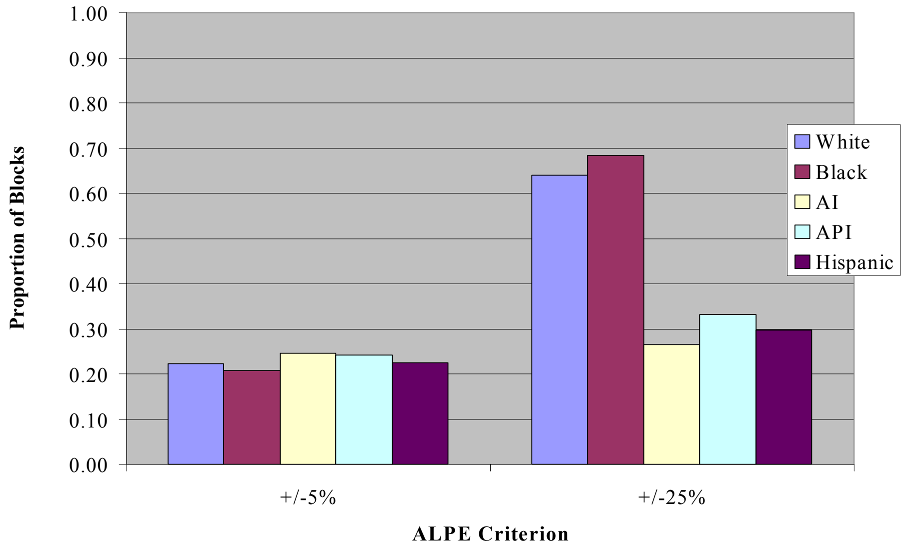


Figure A4.3c: Proportion of Blocks with Race ALPEs Below 5% and 25% - Douglas County

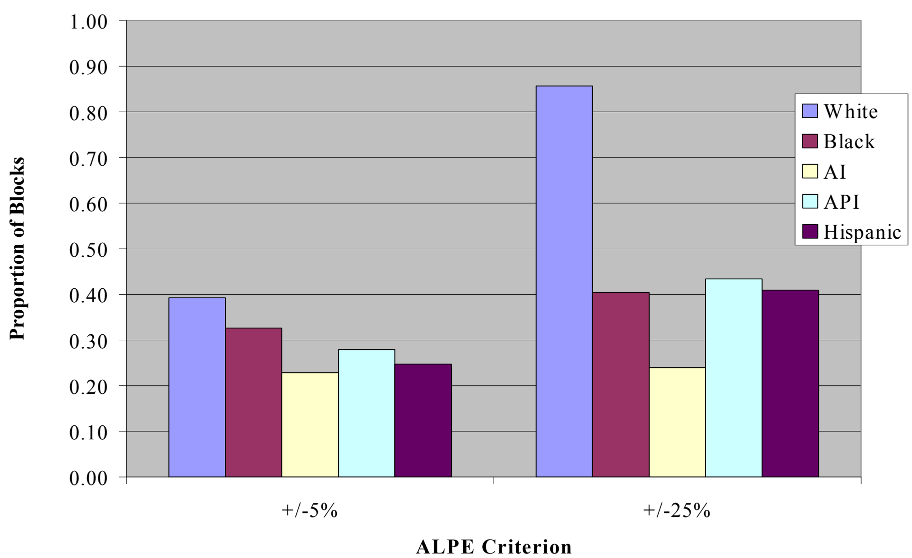


Figure A4.3d: Proportion of Blocks with Race ALPEs Below 5% and 25% - El Paso County

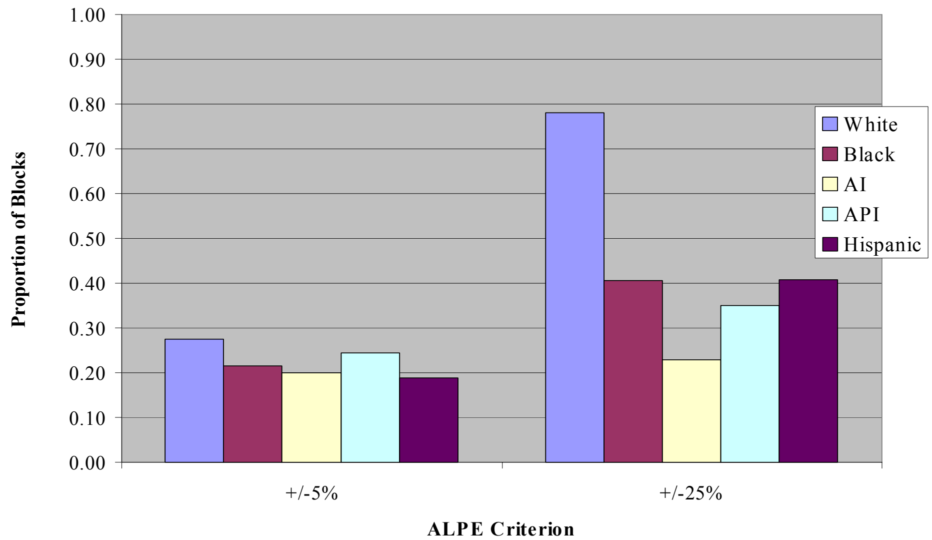
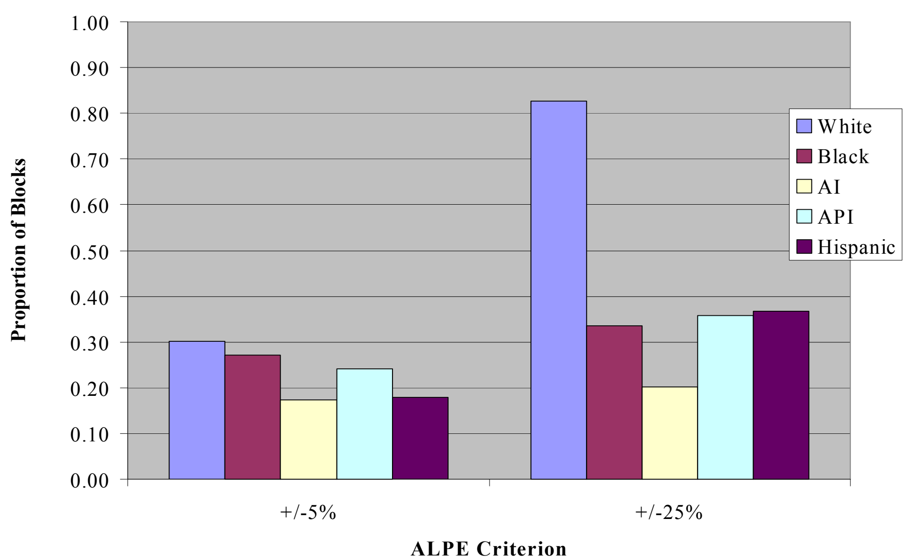


Figure A4.3e: Proportion of Blocks with Race ALPEs Below 5% and 25% - Jefferson County



- In general, ALPE results at the five percent criterion were better for blocks than tracts; but race groups with smaller populations were less accurately counted by AREX.
- All race groups had fewer blocks meeting the 25 percent criterion, compared to tract results.
- In the MD counties, a smaller proportion of blocks were within the five percent criterion for Whites and Blacks, compared to tracts; but a larger proportion of each of the other race groups was within the five percent criterion.

The expected pattern of smaller geography and less accurate AREX counts is supported by the AREX results at the 25 percent criterion. But there is a general tendency for some race groups to be counted more accurately at the block rather than tract-level. This again suggests that cumulative errors may be occurring at tract and county levels, and is especially evident for AIs and APIs.

APPENDIX 5: MULTIVARIATE MODEL PARAMETER ESTIMATES

Table A5.1a: Categorical Logistic Regression Results Predicting Total Block-Level ALPEs-MD (n=13731)[1]

	Large undercount		Moderate undercount		Moderate overcount		Large overcount	
	ß	se	ß	se	ß	se	ß	se
intercept	-1.932 **	0.183	-0.964 **	0.147	-0.923 **	0.153	-2.146 **	0.189
Baltimore City	0.106 **	0.081	-0.165 **	0.064	0.107	0.066	0.722 **	0.086
vacancy rate (> median)	-0.288 **	0.073	-0.060	0.058	0.444 **	0.058	0.811 **	0.073
rental rate (> median)	1.011 **	0.088	0.390 **	0.071	0.050	0.072	0.438 **	0.087
nonrelatives in HH (> median)	0.332 **	0.073	0.174 **	0.058	-0.030	0.057	-0.160 **	0.071
imputed race %-tax (> median)	-0.791 **	0.071	-0.201 **	0.057	0.206 **	0.056	0.173 **	0.070
imputed race %-pcf (> median)	-0.305 **	0.068	-0.010	0.054	0.027	0.053	-0.036	0.067
imputed ethnicity %(> median)	0.851 **	0.124	0.706 **	0.109	0.898 **	0.119	0.880 **	0.133
censpull % (> median)	-0.092	0.077	0.124 **	0.062	0.176 **	0.061	0.080	0.076
multi-race-any mention	-0.268 **	0.073	-0.037	0.057	-0.169 **	0.058	-0.510 **	0.079
some other race-any mention	-0.003	0.091	0.001	0.071	-0.266 **	0.077	-0.215 **	0.113
population density-Q1	0.099	0.108	-0.652 **	0.089	-0.273 **	0.083	0.526 **	0.101
population density-Q2	0.088	0.098	-0.244 **	0.075	-0.175 **	0.074	0.255 **	0.094
population density-Q5	0.162 **	0.082	0.117	0.065	-0.012	0.067	-0.414 **	0.085
neighborhood-1	-0.044	0.138	-0.139	0.109	-0.108	0.105	0.117	0.124
neighborhood-3	0.446 **	0.123	-0.002	0.094	-0.261 **	0.095	-0.033	0.125
neighborhood-4	0.421 **	0.117	-0.020	0.088	-0.130	0.085	-0.039	0.111
White %-Q1	0.243	0.212	-0.714 **	0.199	-0.294	0.207	0.920 **	0.225
White %-Q2	0.096	0.192	-0.619 **	0.181	-0.343	0.194	0.658 **	0.215
White %-Q5	-0.217 **	0.094	-0.175 **	0.074	0.036	0.072	0.422 **	0.094
Blacks-any	-0.358	0.200	0.422 **	0.186	0.633 **	0.198	0.031	0.217
Hispanics-any	-0.316 **	0.080	-0.002	0.061	-0.108	0.061	-0.599 **	0.085
age < 5 (> median)	0.466 **	0.068	0.342 **	0.054	-0.312 **	0.055	-0.596 **	0.071
age 5-19 (> median)	0.149	0.078	0.211 **	0.062	-0.230 **	0.062	-0.564 **	0.078
age 20-24 (> median)	0.398 **	0.070	0.194 **	0.057	0.089	0.057	-0.036	0.071
age 25-44 (> median)	-0.042	0.074	0.173 **	0.059	-0.064	0.059	-0.076	0.073
age 65+ (> median)	-0.265 **	0.077	-0.012	0.062	0.071	0.062	-0.071	0.076

** p < .05; used to distinguish important predictors for these non-sample data.
[1]Reference range is best quartile estimate, where AREX is -2.3% to +5.5% of Census total population

Large undercount < -14.4%; Moderate undercount=-14.4% to -2.3%; Moderate overcount=5.5% to 19.8%; Large overcount > 19.8%

Table A5.1b: Categorical Logistic Regression Results Predicting Total Block-Level ALPEs-CO (n=16948)[1]

	Large undercount		Moderate undercount		Moderate overcount		Large overcount	
	ß	se	ß	se	ß	se	ß	se
intercept	-2.346 **	0.162	-1.707 **	0.139	-1.363 **	0.149	-2.352 **	0.168
Douglas County	-0.126	0.091	0.180 **	0.071	-0.002	0.075	-0.105	0.093
El Paso County	-0.191 **	0.066	-0.253 **	0.055	0.063	0.054	0.043	0.065
vacancy rate (> median)	-0.057	0.088	0.133	0.074	0.550 **	0.074	0.740 **	0.084
rental rate (> median)	0.527 **	0.065	0.238 **	0.054	0.011	0.053	0.193 **	0.065
nonrelatives in HH (> median)	0.456 **	0.064	0.305 **	0.052	0.159 **	0.051	-0.074	0.064
imputed race %-tax (> median)	-0.651 **	0.068	-0.007	0.055	0.355 **	0.054	0.331 **	0.067
imputed race %-pcf (> median)	-0.065	0.064	0.178 **	0.051	0.248 **	0.051	0.521 **	0.064
imputed ethnicity %(> median)	0.936 **	0.099	1.260 **	0.095	1.559 **	0.113	1.393 **	0.113
censpull % (> median)	0.133	0.070	0.342 **	0.058	0.028	0.059	-0.020	0.071
multi-race-any mention	-0.530 **	0.074	-0.219 **	0.056	-0.315 **	0.056	-0.772 **	0.077
some other race-any mention	-0.214 **	0.079	-0.119 **	0.059	-0.182 **	0.059	-0.222 **	0.084
population density-Q1	0.134	0.085	-0.503 **	0.073	-0.318 **	0.070	0.693 **	0.084
population density-Q2	0.172 **	0.077	-0.107	0.061	0.029	0.059	0.468 **	0.078
population density-Q5	0.496 **	0.110	0.299 **	0.089	-0.349 **	0.106	-0.024	0.148
neighborhood-1	0.106	0.111	0.125	0.086	-0.104	0.083	-0.138	0.110
neighborhood-2	0.653 **	0.104	0.170 **	0.086	-0.115	0.086	0.130	0.107
neighborhood-4	0.278 **	0.111	0.030	0.096	-0.302 **	0.092	-0.011	0.102
White %-Q1	0.379	0.332	-2.721 **	1.017	n/a	n/a	0.628 **	0.298
White %-Q2	0.651 **	0.084	-0.026	0.067	-0.095	0.069	0.844 **	0.089
White %-Q5	0.132	0.087	-0.445 **	0.074	-0.445 **	0.072	0.214 **	0.086
Blacks-any	-0.355 **	0.102	-0.130	0.079	0.056	0.080	-0.093	0.105
Hispanics-any	-0.270 **	0.077	0.110	0.065	0.096	0.064	-0.448 **	0.077
age < 5 (> median)	0.537 **	0.067	0.326 **	0.054	-0.291 **	0.055	-0.603 **	0.074
age 5-19 (> median)	0.487 **	0.065	0.271 **	0.055	-0.033	0.055	-0.569 **	0.069
age 20-24 (> median)	0.334 **	0.066	0.149 **	0.055	0.139 **	0.055	0.151 **	0.068
age 25-44 (> median)	-0.145 **	0.065	-0.016	0.054	-0.263 **	0.054	-0.102	0.067
age 65+ (> median)	0.148 **	0.072	0.174 **	0.061	0.183 **	0.059	0.244 **	0.070

** p < .05; used to distinguish important predictors for these non-sample data.

[1]Reference range is best quartile estimate, where AREX is -4.2% to +2.0% of Census total population

Large undercount < -16.7%; Moderate undercount= -16.7% to -4.2%; Moderate overcount=2.0% to 16.2%; Large overcount > 16.2%

Table A5.2a: Piecewise Logistic Regression Results Predicting Total Block-Level ALPEs-MD (n=13731)[1]

			Mobility Variables			Imputation/Process variables			Other Race		Population Density Quintile			Neighborhoods	
Q-grp	Intercept	vac50	rent50	nrel50	imp-tax	imp-pcf	imp-hisp	censpull	mrace	sor	Q1	Q2	Q5	neigh-1	neigh-3
1	-0.289	0.014	-0.016		0.043	0.032		0.026			-0.063	-0.018		-0.027	-0.027
2	-0.068	0.005	-0.004	-0.003	0.004								-0.003		
3	0.110	0.004											-0.005		
4	0.278	0.006			0.005	-0.005			-0.007		0.020	0.012	-0.012		
5	0.010					0.003	0.007	0.003			-0.003			-0.004	-0.005

Table A5.2a (cont'd)

		White Quintiles			Race		County	Age Groups				
Q-grp	neigh-4	Q1	Q2	Q5	Black	Hispanic	Baltimore	< 5	5-19	20-24	25-44	65+
1	-0.029	-0.040	-0.035		0.036		-0.016					0.032
2						0.003				-0.003		0.003
3		0.013				-0.005	0.006					
4		0.021					0.009	-0.010	-0.006			
5	-0.003	-0.006		-0.003	0.007		-0.002					

note: $p < .05$ for all parameter estimates shown; non-significant estimates have been omitted to simplify display; used to distinguish important predictors for these non-sample data.

Q-grps relative to reference group: 1=large undercount, 2=moderate undercount, 3=moderate overcount, 4=large overcount

[1]See notes at end of Table A5.2f for additional information.

Table A5.2b: Piecewise Logistic Regression Results Predicting Total Block-Level ALPEs-CO (n=16948)[1]

			Mobility Variables			Imputation/Process variables			Other Race		Population Density Quintile			Neighborhoods	
Q-grp	Intercept	vac50	rent50	nrel50	imp-tax	imp-pcf	imp-hisp	censpull	mrace	sor	Q1	Q2	Q5	neigh-1	neigh-2
1	-0.301			0.015	0.056	0.032		0.019			-0.053	-0.030	-0.020		-0.034
2	-0.092		-0.003		0.004								-0.006		-0.005
3	0.089								-0.004	-0.003	0.008			-0.005	
4	0.258			-0.007			-0.013		-0.007		0.022				0.011
5				-0.002			-0.006	-0.002			0.002				

Table A5.2b (cont'd)

		White Quintiles			Race		County	Age Groups					
Q-grp	neigh-4	Q1	Q2	Q5	Black	Hispanic	Douglas	El Paso	< 5	5-19	20-24	25-44	65+
1	-0.026	-0.100	-0.032		0.024	0.015			-0.014	-0.016		0.013	
2	-0.005		-0.006						-0.003				
3			0.008			-0.005			-0.004	-0.005			
4	0.014	0.052	0.026	0.020		-0.012			-0.007	-0.012	-0.009		
5				0.003		-0.003			-0.002	-0.002			

note: $p < .05$ for all parameter estimates shown; non-significant estimates have been omitted to simplify display; used to distinguish important predictors for these non-sample data.

Q-grps relative to reference group: 1=large undercount, 2=moderate undercount, 3=moderate overcount, 4=large overcount

[1]See notes at end of Table A5.2f for additional information.

Table A5.2c: Piecewise Logistic Regression Results Predicting Age 0-4 Block-Level ALPEs-CO (n=12603 for all 5 models) [1]

		Mobility Variables			Imputation/Process variables				Other Race		Population Density Quintile			Neighborhoods	
Q-grp	Intercept	vac50	rent50	nrel50	imp-tax	imp-pcf	imp-hisp	censpull	mrace	sor	Q1	Q2	Q5	neigh-1	neigh-2
1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
2	-0.543		-0.011		0.026		-0.034	0.010							
3	-0.319				0.013		-0.017	0.013							
4	0.754								-0.163						
5	0.037				-0.011		-0.031	-0.013	-0.022	-0.013	0.008	0.006			

Table A5.2c (cont'd)

		White Quintiles			Race		County	Age Groups					
Q-grp	neigh-4	Q1	Q2	Q5	Black	Hispanic	Douglas	El Paso	< 5	5-19	20-24	25-44	65+
1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
2				0.013		-0.017							0.014
3				0.010					-0.014	-0.006			
4			0.112		-0.131				-0.430				0.172
5			0.023	0.006	-0.015	-0.018	-0.012		-0.039				0.008

note: $p < .05$ for all parameter estimates shown; non-significant estimates have been omitted to simplify display; used to distinguish important predictors for these non-sample data.

Q-grps relative to reference group: 1=large undercount, 2=moderate undercount, 3=moderate overcount, 4=large overcount

[1]See notes at end of Table A5.2f for additional information.

Table A5.2d: Piecewise Logistic Regression Results Predicting Age 65+ Block-Level ALPEs-MD (n=12688 for all 5 models) [1]

			Mobility Variables		Imputation/Process variables				Other Race		Population Density Quintile			Neighborhoods	
Q-grp	Intercept	vac50	rent50	nrel50	imp-tax	imp-pcf	imp-hisp	censpull	mrace	sor	Q1	Q2	Q5	neigh-1	neigh-3
1	-0.534	0.039			0.029	0.067	0.063	0.048	0.044		-0.155	-0.076		-0.054	-0.091
2	0.098		-0.007						-0.006				-0.005		
3	0.275	0.008													
4	0.741										0.289	0.190			0.193
5							-0.008		-0.005	-0.007	0.010		-0.007		0.007

Table A5.2d (cont'd)

		White Quintiles			Race		County	Age Groups				
Q-grp	neigh-4	Q1	Q2	Q5	Black	Hispanic	Baltimore	< 5	5-19	20-24	25-44	65+
1	-0.081	-0.164	-0.137		0.110	0.056			0.031			0.107
2						-0.007	0.008					-0.012
3						-0.007		0.007				-0.020
4		0.271										-0.281
5	0.006			0.004		-0.007					0.004	-0.015

note: $p < .05$ for all parameter estimates shown; non-significant estimates have been omitted to simplify display; used to distinguish important predictors for these non-sample data.

Q-grps relative to reference group: 1=large undercount, 2=moderate undercount, 3=moderate overcount, 4=large overcount

[1] See notes at end of Table A5.2f for additional information.

Table A5.2e: Piecewise Logistic Regression Results Predicting Block-Level Black ALPEs-MD (n=11238 for all 5 models)[1]

			Mobility Variables			Imputation/Process variables			Other Race		Population Density Quintile			Neighborhoods	
Q-grp	Intercept	vac50	rent50	nrel50	imp-tax	imp-pcf	imp-hisp	censpull	mrace	sor	Q1	Q2	Q5	neigh-1	neigh-3
1	-0.778					0.048		0.045			-0.069				
2	-0.254		-0.018		0.012				0.021	0.012	-0.020				
3	0.173							-0.009			0.017		-0.013		
4	0.952			0.330	-0.303			-0.323			0.910				
5															

Table A5.2e (cont'd)

		White Quintiles			Race		County	Age Groups				
Q-grp	neigh-4	Q1	Q2	Q5	Black	Hispanic	Baltimore	< 5	5-19	20-24	25-44	65+
1	-0.046			-0.140		0.042		0.029		0.031		-0.031
2		0.047	0.038	-0.123	0.027	0.011		-0.008				
3					-0.030	-0.011	0.017	-0.007				
4												
5												

note: $p < .05$ for all parameter estimates shown; non-significant estimates have been omitted to simplify display; used to distinguish important predictors for these non-sample data.

Q-grps relative to reference group: 1=large undercount, 2=moderate undercount, 3=moderate overcount, 4=large overcount

[1]See notes at end of Table A5.2f for additional information.

Table A5.2f: Piecewise Logistic Regression Results Predicting Block-Level Hispanic ALPEs-CO (n=11289 for all 5 models) [1]

			Mobility Variables			Imputation/Process variables			Other Race		Population Density Quintile			Neighborhoods	
Q-grp	Intercept	vac50	rent50	nrel50	imp-tax	imp-pcf	imp-hisp	censpull	mrace	sor	Q1	Q2	Q5	neigh-1	neigh-2
1	-0.989		0.025	0.013	0.013	0.047			0.030	0.041	-0.026	-0.020	-0.028		
2	-0.472				0.016	0.054		0.014	0.022	0.012	-0.025	-0.016	-0.023	0.019	
3	0.434								-0.047	-0.071			-0.039		
4	1.494				0.336	0.210				-0.303					
5	0.017		-0.009		-0.012	-0.021	-0.020		-0.019	-0.029	0.009				

Table A5.2f (cont'd)

			White Quintiles		Race		County	Age Groups					
Q-grp	neigh-4	Q1	Q2	Q5	Black	Hispanic	Douglas	El Paso	< 5	5-19	20-24	25-44	65+
1					0.026					0.017			
2			-0.017		0.022			-0.014					
3													
4				-0.388									
5							0.009	0.006			-0.006		

note: p < .05 for all parameter estimates shown; non-significant estimates have been omitted to simplify display; used to distinguish important predictors for these non-sample data.

Q-grps relative to reference group: 1=large undercount, 2=moderate undercount, 3=moderate overcount, 4=large overcount

vac50=vacant proportion of block housing units: binary indicator of top 50%
rent50=rental proportion of block housing units: binary indicator of top 50%
nrel50=non-relative household member proportion of block households: binary indicator of top 50%
imp-tax=proportion of cases imputed using tax method: binary indicator of top 50%
imp-pcf=proportion of cases imputed using pcf method: binary indicator of top 50%
imp-hisp=proportion of cases imputed for hispanic origin: binary indicator of top 50%
censpull=proportion of census pull cases: binary indicator of top 50%
mrace=presence of multi-race reports

sor=presence of some other race reports
popd Q1, Q2, Q5=population density quintiles of block: Q1 (low), Q2, Q5 with Q3 and Q4 as the reference group
neigh-1, 2, 3, 4=type of neighborhood from factor analyses
wht Q1, Q2, Q5=White quintile groups from White proportion of block population: Q1 (low), Q2, Q5 with Q3 and Q4 as the reference group
Black=Black proportion of block population: binary indicator of top 50%
hispanic=presence of any Hispanic residents on block
baltimore=indicator for Baltimore City: Baltimore County as reference group; douglas, elpaso=indicators for counties: Jefferson County as reference group
age < 5=age 0-4 proportion of block population: binary indicator of top 50%
age 5-19=age 5-19 proportion of block population: binary indicator of top 50%
age 20-24=age 0-4 proportion of block population: binary indicator of top 50%
age 25-44=age 25-44 proportion of block population: binary indicator of top 50%
age 65+=age 65+ proportion of block population: binary indicator of top 50%

APPENDIX 6: GLOSSARY OF TERMS AND DEFINITIONS

ALPE	Algebraic Percent Error, formed from Census and AREX counts using Census results as the standard.
ABI	American Business Information; ABI is a commercially available list of residential and business addresses covering the entire U.S.
AI	American Indians.
API	Asian and Pacific Islanders.
AREX 2000	Administrative Records Experiment in 2000.
Bottom-up	Bottom-up method of processing AREX counts that includes MAF address verification and variable imputation.
Census-pull	For addresses that failed to match the MAF, the bottom-up process replaced some of these addresses with actual Census 2000 records.
Code-1	Code-1 is a commercially available software product used to standardize and match addresses to other address lists.
FAV estimation	For addresses that failed to match the MAF, the bottom-up process replaced some of these addresses using estimated counts derived from a sample of households that were authenticated by a field address verification (FAV) process.
GIS	Geographic information system.
Hispanic origin	Hispanic origin of any type, based on administrative reports, surname processing, country of origin, and Hispanic origin of householder.
Hot deck assignment	The race imputation process used statistical models to calculate expected race probabilities for each person. The hot deck assignment was based on an algorithm that compared the calculated probability with a randomly drawn number to determine whether a calculated probability was large enough to be assigned to a particular race category.
Index of Dissimilarity	Index of summed differences between AREX and Census counts based on either race/ethnicity or age groups.
MAF or Master Address File	The master list of verified household addresses used to conduct Census-related activities.
Multi-race rate	Derived from Census: based on reported number of race responses.
Neighborhood characteristics	Estimated from factor analyses that distinguish four types of AREX neighborhoods in each AREX state; derived from demographic, housing unit, and population density variables.

Non-relative rate	Derived from Census: proportion of households with non-relative members.
NRFU	Nonresponse follow-up; households that could not be enumerated through usual Census enumeration methods.
Numident	The electronic roster of participants in any of the social programs maintained by the Social Security Administration, compiled from SSN applications, name changes, and corrections.
Overcount	AREX counts that are greater than Census counts, expressed as differences or ALPEs.
PCF probability model	The personal characteristics file (PCF) used a probabilistic race imputation methodology based on logistic regression models and hot deck assignment.
Population density	Population per unit area, expressed as persons per square mile.
PRED	Planning, Research, and Evaluation Division.
Race	AREX race values are based on 'generally accepted' race categories that are derived from complex AREX processing rules; Census race measures use self-reported race from Census forms and exclude persons claiming some other race or multi-race.
Rental rate	Derived from Census: proportion of housing units identified as rental units.
Shannon-Wiener Index of Diversity	Summed index of age or race components using AREX-only measures to distinguish regions with more or less diverse populations.
StARS	Statistical Administrative Records System.
Top-down	Top-Down Administrative Records counts that includes block-coding but no further enhancements.
TIGER	Topologically Integrated Geographic and Cartographic Encoding and Referencing database of all U.S. regions and Puerto Rico.
Undercount	AREX counts that are less than Census counts, expressed as differences or ALPEs.
Vacancy rate	Derived from Census: proportion of housing units identified as vacant.